A Computational Framework for Modeling Belief-based Decision Making

Koosha Khalvati

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Rajesh P. N. Rao, Chair

Adrienne Fairhall

Andrea Stocco

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

A Computational Framework for Modeling Belief-based Decision Making

Koosha Khalvati

Chair of the Supervisory Committee:
Professor Rajesh P. N. Rao
Computer Science and Engineering

Existing computational models of decision making are often limited to particular experimental setups. The limitation is mainly due to the inability to capture the decision maker's uncertainty about the situation. We propose a computational framework for studying decision making under uncertainty in neuroscience and psychology. Our framework is heavily focused on the probabilistic assessment of the decision maker, i.e., their "belief", about the state of the world. Specifically, it is based on Partially Observable Markov Decision Processes (POMDPs), which combines Bayesian reasoning and reward maximization to choose actions. We demonstrate the viability of our belief-based decision making framework using data from various experiments in perceptual and social decision making. Our framework explains the relationship between decision makers' actual performance and their belief about it, called decision confidence, in perceptual decision making experiments. It also shows why this assessment could deviate from reality in many situations. Such deviations have been often interpreted as evidence for sub-optimal decision making or distinct processes that underlie choice and confidence. Our framework challenges these interpretations by showing that a normative Bayesian decision maker optimizing the gained reward elicits the same discrepancies. Moreover, our method outperforms existing models in quantitatively predicting human behavior in a social decision making task and provides insight into the underlying process. Our results suggest that in decision making tasks involving large groups, humans employ Bayesian inference to model the "group's mind" and make predictions of others' decisions. Finally, we ex-

tend our method to multiple reasoning levels about others (levels of theory of mind) and make the connection to conformity as a strategy for decision making in groups. This extended framework can explain human actions in various collective group decision making tasks, providing a new theory for cooperation and coordination in large groups.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I first and foremost express my sincere appreciation to my supervisor Rajesh Rao for his guidance and support during my PhD. He is one of the most caring advisers one could have. I like to deeply thank Roozbeh Kiani for having me in his lab at NYU during our collaboration in one of my PhD projects. I also thank Jean-Claude Dreher for our collaboration in social decision making projects. I extremely appreciate the help and guidance of my committee members Eric Shea-Brown, Adrienne Fairhall, Andrea Stocco, and Daniel Weld.

Success of my projects is owed to collaboration and discussion with members of Neural Systems Lab at the University of Washington, Kiani Lab at New York University, and Neuroeconomics Lab at the Institute of Cognitive Sciences Marc Jeannerod. I am greatly thankful to all of my collaborators and colleagues during my PhD especially Saghar Mirbagheri, Seongmin Park, and Dimitrios Gklezakos. I also thank Shinsuke Suzuki and John O'Doherty for sharing their data with me for one of my projects. Faculty and staffs of Allen school played a huge role in my academic success by creating a welcoming and supporting environment. I especially appreciate the extraordinary job that Elise deGoede Dorough and Lisa Merlin are doing in our department.

Graduate school was a very joyful experience for me because of many great friends at the Allen school Hessam Bagherinezhad, Alireza Rezaei, Ignacio Cano, Deepali Aneja, Srinivasan Iyer, and many others. I am also deeply grateful to my friends outside of school especially Amir Mehrabian, Chakaveh Ahmadizadeh, Nazli Akhtari, and Shaham Shafiee Fazel.

I would be always thankful to my family Kafieh Naziri, Mohammad Ali Khalvati, and Roxana Niktab for their unconditional and eternal love and support. Finally, I want to express my deepest gratitude to my love Saghar Mirbagheri for her constant emotional support in addition to her tremendous help and contribution toward my academic growth.

Chapter 1

INTRODUCTION

Decision making has been the subject of interest in many neuroscience and psychology studies for many years [126, 91, 64]. Most developed computational models of decision making, however, are still minimal. These models can explain humans and non-verbal animals' behavior in particular experiments. However, they are often not generalizable to other tasks and, most importantly, to situations that animals face daily. This limitation is primarily due to the inability of the existing methods in modeling uncertainty and the decision maker's assessment of the situation based on noisy and ambiguous inputs.

For example, practiced reinforcement learning methods based on Markov Decision Processes (MDPs) and temporal difference (TD) learning [143] are very successful in explaining animals' behavior in experimental tasks [9, 33, 95]. In these experiments, the state of the environment is fully observable to the decision maker. However, the real world is always only partially observable to the decision maker due to its complexity and the limits of the decision maker's sensory system.

How animals deal with the ambiguity of the state of the world, i.e., perceptual ambiguity has been extensively studied in neuroscience, in a field named perceptual decision making[64]. However, almost all of the field's computational models have ignored the role of actions, rewards, and most importantly, the decision maker's internal model of the task. This internal model plays a crucial role in the decision maker's assessment of the environment's state.

Here we utilize a method developed in Artificial Intelligence (AI), Partially Observable Markov Decision Process (POMDP), to model decision making under uncertainty in neuroscience studies. This Reinforcement Learning (RL) framework embraces various aspects of decision making in the real world, including reasoning under uncertainty and utility maximization. Figure 1.1 shows a schematic of this framework where the agent interacts with the environment through actions,

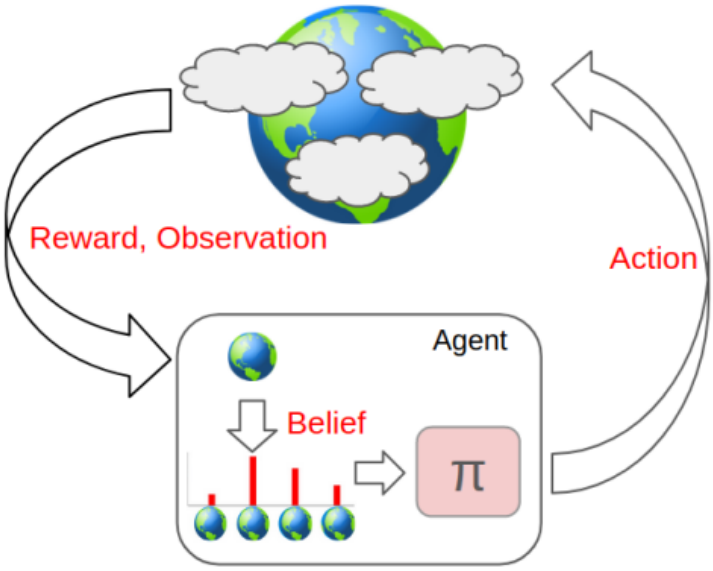


Figure 1.1: **Belief-based decision making framework**. The agent interacts with the world through actions, observations, and reward. The state of the world is not fully observable to the agent and is represented probabilistically based on observations and agent's internal model of the world. The goal of the agent is to come up with a policy based on the probability distribution over the current state, known as the *belief*.

observations, and reward. Specifically, the agent receives reward and observation after each action, depending on the environment's current state. Moreover, the actions of the agent can change the state of the world.

The agent's goal is to come up with a recipe for action selection, called *policy*, that maximizes total reward (utility) in the long run. The state of the world is not fully available. However, the decision maker's gained reward depends on it. Therefore, the agent needs to develop a *belief* about the state. This belief is a probabilistic representation of the current state, based on received observations and the agent's (internal) world model. It plays a crucial role in finding a good policy to obtain more reward.

Three aspects of our *belief-based* framework make it especially useful for computational mod-

eling of the brain and behavior:

- **Applicability to real-world situation:** Since POMDPs can handle uncertainty, they can be applied to real-world scenarios where uncertainty and ambiguity always exist. The most common sources of uncertainty are the limitations of the sensory system, the existence of noise in sensory and motor systems, and lack of full knowledge about the surrounding environment. Our belief-based framework can model uncertainty caused by any of these sources.

- **Generalizability:** Our framework is applicable to different situations and experiments involving sensory, motor, or reward system. Moreover, it can model any task once its components such as state space, action space, and observations are identified. While identifying these components is sometimes a research question in itself, the rest of modeling would be straightforward after this step.

- **Founded on principles of computation:** The belief-based framework suggests action selection based on some principles, including existence of an internal model, Bayesian update of the belief, and reward maximization. Therefore, modeling animals' behavior with this framework gives us insight into the brain's computational mechanisms. It also suggests experimental designs for future studies of decision making to test different hypotheses.

1.1 Organization

In this thesis, we test our framework on perceptual and social decision making experiments and answer scientific questions based on our modeling results in both domains. These tasks are very different from each other in many aspects, but our framework can model them due to its generalizability and flexibility. Moreover, in both types of tasks, the subject's belief about the unobservables plays a considerable role. Perceptual decision making is about finding the hidden state of the environment based on noisy inputs. Also, in social decision making, one should choose actions considering others' intentions and reactions, both unavailable when deciding.

Chapter 2

We start with a review of common computational approaches for decision making in neuroscience and AI in chapter 2. This chapter's primary focus is on methods related to our belief-based framework, including reinforcement learning models and algorithms, and decision theories of perceptual decision making. We continue by giving the formal definition of the POMDP framework in AI and the main challenges in solving it. Finally, we explain why the POMDP framework alone is insufficient for analyzing animal behavior and what should be considered in applying POMDP to neuroscience experiments.

Chapter 3

Chapter 3 is about perceptual decision making. It starts with modeling perceptual decision making tasks with our belief-based framework. Our approach models the subject's choice and their belief about it in a unifying framework. This approach enables us to predict subject's belief about their choice, i.e., their decision confidence, solely from performance in perceptual decision making experiments. We test our model on monkeys performing a direction-discrimination task with post-decision wagering [86]. Finally, we show that our model can be implemented by simple mechanisms that accumulate evidence toward a bound.

Chapter 4

In chapter 4, we show that our framework explains several commonly observed phenomena related to the relationship between confidence and performance. These observations include the hard-easy effect [40, 133], opposing effects of stimulus variability on confidence and accuracy [169], dependence of confidence ratings on simultaneous or sequential reports of choice and confidence [79, 87], apparent difference between choice and confidence sensitivity [92, 48], and seemingly disproportionate influence of choice-congruent evidence on confidence [168, 114]. These phenomena have been mostly interpreted as evidence for suboptimality of decision making or separate mechanisms behind the belief and choice in the brain. However, our results show that they arise in

Bayesian inference with incomplete knowledge of the environment.

Chapter 5

Chapter 5 is about social decision making, in which we show our belief-based framework can also very well explain human behavior when interacting with others. To make proper decisions in a social context, humans have to predict the behavior of others [146]. This ability relies on having a model of other minds known as theory of mind (ToM). Such a model becomes especially complex in group decision making where the number of people one simultaneously interacts with is large. We create a model based on our belief-based framework and test it on a group decision making task known as the Volunteer's Dilemma (VD). Our Bayesian belief-based model outperforms existing models in quantitatively predicting behavior and different outcomes of group interactions. Our results suggest that in group decision making tasks, humans use Bayesian inference to model an average group member's mind or "mind of the group". Based on this built model, they make predictions of others' decisions while also simulating the effects of their own actions on the group as a whole in the future.

Chapter 6

In chapter 6, which is also about social decision making, we show how one can incorporate different levels of "sophistication" into our belief-based framework and explain a wide range of collective decision making tasks among groups. In collective decision making, members of a group need to coordinate their actions to achieve a desirable outcome. Collective decision making could be extremely challenging when there is some form of competition between group members, when the group is large, and most importantly when there is no direct communication between the players. In this chapter, we present a new Bayesian theory of collective decision making based on a simple yet most commonly observed behavior, i.e., conformity. Specifically, using our belief-based model, we show how humans are extremely good at collective decision making by utilizing the multi-level Bayesian Theory of Mind based on conformity. Our results include quantitative fits to data from

three different experiments with around 300 subjects in total. Not only can our framework quantitatively explain human behavior in various group decision making tasks, but the predicted levels of ToM change meaningfully both in each task and between different tasks. Our results suggest that conformity, reasoning about the mind of the group as a whole, and simulation of future events are the bases of collective decision making in groups.

Chapter 7

Finally, chapter 7 draws the connection between previous chapters by summarizing them through our belief-based framework. We also discuss the main contributions of this thesis. Moreover, we explain how the framework can help us with better experimental design and with testing different hypotheses about how the brain makes decisions in the face of uncertainty. Future directions of this framework is the last part of the last chapter.

Chapter 2

COMPUTATIONAL MODELS OF DECISION MAKING IN NEUROSCIENCE AND AI

The practice of using methods rooted in Artificial Intelligence (AI) to analyze neural and behavioral data has been significantly increased in recent years [164, 14, 13, 60]. Achieving near human-level performance in various tasks, being founded on generalizable rules and principles, transparency and adjustability of every single component, and the emergence of similar phenomena observed in neuroscience studies are the major reasons for this increase. Reinforcement learning (RL) methods are not an exception. In fact, reinforcement learning has been always an area of interest for both communities of Artificial Intelligence and neuroscience [143, 60].

Reinforcement leaning is concerned about developing algorithms for an agent that seeks to maximize its total reward by interacting with its environment [143]. As a result, it is naturally a suitable framework to study decision making. Practiced RL methods in neuroscience, however, mostly ignore the presence of uncertainty. This is a major problem for explaining real life decisions as uncertainty is always present in the real world. On the other hand, dealing with uncertainty is not easy. Reinforcement learning methods that can handle uncertainty are complex and computationally very expensive. In fact, even in the field of Artificial Intelligence, methods that deal with uncertainty are not commonly used.

We start this chapter by reviewing commonly used RL approaches in AI and neuroscience. Then, we talk about computational methods that model uncertainty in neuroscience. After that, we explain Partially Observable Markov Decision Process (POMDP), an RL method for decision making under uncertainty in AI [75], followed by the main challenge in using it. Finally, we explain how to adapt POMDPs to be applicable in neuroscience and psychology studies.

2.1 Reinforcement learning models

A reinforcement learning problem is usually expressed as a Markov Decision Process (MDP), describing all possible states of the environment, its dynamics, and how the agent interacts with it. Formally, MDP is (S, A, T, R, γ) with following definitions. S is the set of states of the environment. A is the set of all available actions to the agent. Transition function $T : |S| \times |A| \times |S| \to [0, 1]$ defines $T(s, a, s') = P(s'|s, a)$, the probability of ending up in state s' by performing action a in state s. The Markovian nature of MDP comes from the transition function, i.e. the next state only depends on the current state and current action. $R : |S| \times |A| \to \mathbb{R}$ is a bounded function determining the reward gained in state s by performing action a, showed as $R(s, a)$. Finally, $\gamma \in (0, 1]$ is the discount factor for reward [150].

Starting from initial state, s_0, the goal of an RL agent is to come up with a recipe for action selection, called policy π, that maximizes the total expected reward. Since the system is Markovian, policy could be expressed as a mapping from states to a distribution of actions, i.e. $\pi : |S| \times |A| \to [0, 1]$. The optimal policy π^* is defined as following:

$$\pi^* = \arg\max_{\pi} \sum_{t=0}^{H} \gamma^t \mathbb{E}[R(s_t, a_t)|\pi, s_0]. \tag{2.1}$$

Horizon H defines the length of this sequence and could be infinite. In the case of infinite horizon, the discount factor must be less than 1. Importantly, for any MDP there exists a deterministic optimal policy, which could be expressed as a mapping from states to actions, i.e. $|S| \to |A|$ [128].

In developing algorithms for finding the optimal policy, value and Q function of a policy, i.e. $V^\pi : |S| \to \mathbb{R}$ and $Q^\pi : |S| \times |A| \to \mathbb{R}$ are also defined. $V^\pi(s)$ is simply the expected discounted reward of policy π starting from state s. $Q^\pi(s, a)$ is also the expected discounted reward of policy π if the agent starts from state s and performs action a as the first action.

2.1.1 Reinforcement learning algorithms for finding the optimal policy

Algorithms for finding the optimal policy of an MDP are generally divided into two categories: "model-free" and "model-based". Model-based approaches use the structure of the environment,

i.e transition and reward function to determine the optimal policy (figure 2.1a). Value iteration algorithm is one of the most popular approaches in model-based reinforcement learning. In this method, starting from $V_0(s) = 0$, the value function gets updated with Bellman equation iteratively [12]:

$$V_{k+1}(s) = \max_a \left[R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') V_k(s') \right] \tag{2.2}$$

In a finite horizon MDP, H iterations are needed and $V_i(s)$ is the value of state s when i steps are remained. In the infinite case, $V_k(s)$ is guaranteed to converge to the optimal value function V^* in finite number of iterations [150]. Moreover, the optimal policy can be obtained from optimal value functions in a very similar way to Bellman equation:

$$\pi^*(s) = \arg\max_a \left[R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') V^*(s') \right] \tag{2.3}$$

Other well-known model-based methods are policy iteration, which is very similar to value iteration, and Monte Carlo search. The latter is especially interesting to neuroscientists as it is analogous to *mental simulation* in the brain. In this approach, the agent simulates consequences of each action based on the learned model to come up with the best strategy [60]. Model-based approaches are computationally expensive. For example, value iteration algorithm consist of many iterations, each with complexity of $O(|S|^2|A|)$. These methods, however, are very flexible. With a change in environment, e.g. in transition or reward functions, the agent recalculates the optimal policy based on the updated component.

In the model-free approach, actions are chosen directly based on their reward history without considering the structure of the task (figure 2.1b). For example, Q-learning, a Temporal Difference (TD) algorithm is a model-free method based on Bellman equation [71]. In this approach, starting from an initial value, Q function is updated as following after performing action a_t in state s_t, receiving r_{t+1} as the reward, and ending up in state s_{t+1}:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha_t(s_t,a_t) \left(r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t) \right) \tag{2.4}$$

The learning rate $\alpha_t(s,a)$ is between 0 and 1 and defines the weight that the agent gives to its most recent experience, compared to previous ones. Therefore, learning rate should gradually

decrease as the agent collects more samples. In fact, Q values converge to their optimum value $Q^*(s,a)$ [71], if the state and action spaces are finite, and:

$$\forall(s,a) \in S \times A : \sum_t \alpha_t(s,a) = \infty \quad \text{and} \quad \sum_t \alpha_t^2(s,a) < \infty \tag{2.5}$$

The optimal policy can be easily obtained from optimal Q functions by choosing the action with maximum Q function in each state. Another well-known model-free reinforcement learning algorithm is SARSA (state-action-reward–state–action), which is basically the same as Q-learning with one small difference [129]. In the update rule, SARSA uses the Q value of the next state and next action, i.e. $Q(s_{t+1}, a_{t+1})$ instead of maximum Q value of the next state ($max_a Q(s_{t+1}, a)$). Model-free approaches are computationally very cheap, e.g. one simple update rule, but not flexible. Without knowledge of transition and reward functions, a small change in the environment makes the agent gather all the data again to find out the optimal policy.

2.1.2 Reinforcement learning methods in neuroscience

Reinforcement learning and specifically model-based versus model-free approach has been the subject of many studies in neuroscience in recent years [33, 95, 60, 26]. As mentioned, model-based approaches are very flexible and can adapt very quickly in case of any change in the environment. This flexibility and accuracy, however, comes with a huge computational cost. On the other hand, computationally cheap model-free learning approach might perform very poorly before gaining a lot of experience. Based on the cognitive resources, complexity and importance of the task, and gained utility of different choices the brain chooses one of the strategies or a mixture of them [26].

While these experiments are very useful in studying neuropsychiatric disorders or the role of different brain regions in decision making [26], they cannot be generalized into real-world situations. Due to limits and noise in the sensory system, the outside world is always only partially observable. In other words, the current state of the environment is not fully available to the decision maker. This violates the basic assumptions of Markov decision processes, and consequently model-based and model-free algorithms developed for it.

When the state of the environment is not fully observable, in order to gain reasonably high utility, the decision maker should take the uncertainty about the state of the world into account. This means that instead of "the current state of the environment," the decision maker should consider multiple possible states, some more probable than others. In other words, instead of knowing the current state, the decision maker holds a *belief* about it (Fig. 2.1c). In this case the decision maker performs "belief-based" decision making [51]. Belief-based decision making is a model-based method. However, due to the key role of the belief, and ignoring the uncertainty in most experiments that study model-based versus model-free RL, we separate it from model-based methods. Moreover, dichotomy of model-based versus model-free MDP faces several issues in neuroscience and psychology studies [26]. For example, very similar behavior could emerge from these approaches in many experiments [30]. In fact, because of these issues and persistent existence of uncertainty in the real world, studying behavior as "belief-free" versus "belief-based" strategy might be more fruitful [51].

Belief-based reinforcement learning can be formally defined and solved by Partially Observable Markov Decision Processes (POMDPs) in the filed of Artificial Intelligence [75]. However, it has not been widely used in the neuroscience community yet. While dealing with uncertainty has been studied extensively in neuroscience, almost none of the existing approaches consider the role of decision maker's actions and the reward (or cost) into account. In other words, these studies are mostly limited to perception. In fact, they are widely used in a sub-filed of neuroscience named perceptual decision making. Before discussing belief-based Reinforcement learning and POMDP, we review computational models of perceptual decision making.

2.2 Computational models of perceptual decision making

In perceptual decisions, the subject infers the hidden state of the environment based on noisy sensory information to obtain reward. In most studies the decision maker chooses between two options, e.g. decides whether a noisy image contains a face or a house. In some experiments, especially recently, the subjects also report their belief about the chance of success, called confidence [116], at the end of the trial. (Fig 2.2a).

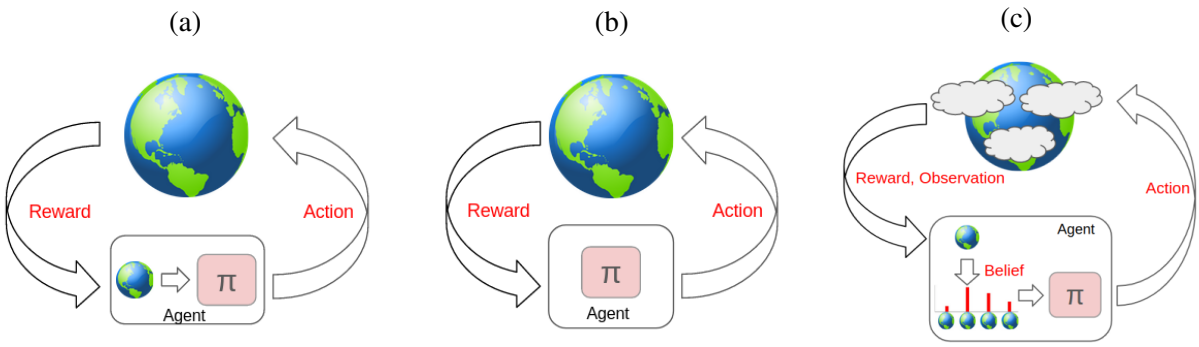


Figure 2.1: **Reinforcement learning methods for decision making** a) In model-free reinforcement learning the agent chooses the action based on the history of rewards directly. This means that it does not consider the structure of the task. b) Model-based decision maker's strategy is to choose the action based on stimulating future events starting from the current state. This simulation is done by using the task structure, especially the transition function which defines how actions change the state of the world. c) In belief-based decision making, the agent does not have a full access to the state of the world. As a result, it simulates the future events based on the probability distribution over the current state.

2.2.1 Drift diffusion model

One of the most common approaches in perceptual decision making is the Drift Diffusion Model (DDM) [123]. DDM assumes that each observation confers evidence in favor of one choice and an equal amount of evidence against the other choice. Integration of sensory evidence over time provides a decision variable (DV) that tracks the total evidence in favor of each choice. In other words, at each time point the decision variable is simply the sum of all pieces of sensory evidence gathered so far. DDM also assumes that each momentary sensory evidence is drawn from a Gaussian distribution. In most formulations of DDM, two constant bounds above and below the initial value of the DV act as termination criteria for the decision. These constant bounds guarantee an

average accuracy with minimum amount of evidence according to Sequential Probability Ratio Test (SPRT) [158]. As soon as the DV reaches one of these bounds, the decision-making process stops and the choice associated with the bound is made. In cases where the stimulus terminates before a bound is reached, the choice with the most supporting evidence is selected (Fig. 2.2b).

DDM has been very successful in explaining subjects' choice accuracy based on stimulus difficulty (e.g. amount of noise) and duration. It can also explain the total amount of time the decision maker needs to respond, called reaction time, if the bounds collapse with time [22]. Collapsing bounds of DDM, however, are not derived in a principled normative way. They are derived from fits with ad hoc urgency signals [22]. Moreover, DDM does not model the belief of the subject about the success (confidence). Therefore, when modeling decision making with DDM, confidence report (or a behavior based on it) should be recorded, and *fitted* based on the accumulated evidence and time [86]. In other words, DDM does not naturally model the belief of the decision maker.

2.2.2 Signal detection theory

Another well-known approach used in perceptual decision making is Signal Detection Theory (SDT) [104] in which the decision is made based on distribution of sensory inputs generated from each hidden state. SDT could be interpreted as a Bayesian approach where the decision maker has the generative model of inputs and computes the probability of each state upon receiving an input. Then, the state with higher probability would be selected. This reasoning could be mapped to a simple decision criterion. Since SDT is Bayesian, decision maker's belief (confidence) would be naturally modeled as well. While the choice depends on which side of the decision criterion the input falls, confidence depends on the distance between input and the criterion [79]. Moreover, discrete confidence report such low vs. high could be mapped to "confidence criterion" [92, 114] (Fig. 2.2c).

The major problem with SDT is that instead of considering each trial as a sequence of observations over time, it looks at all observations all together. This is problematic when decision making takes a different amount time in each trial, either because the time of each trial is in subject's hand

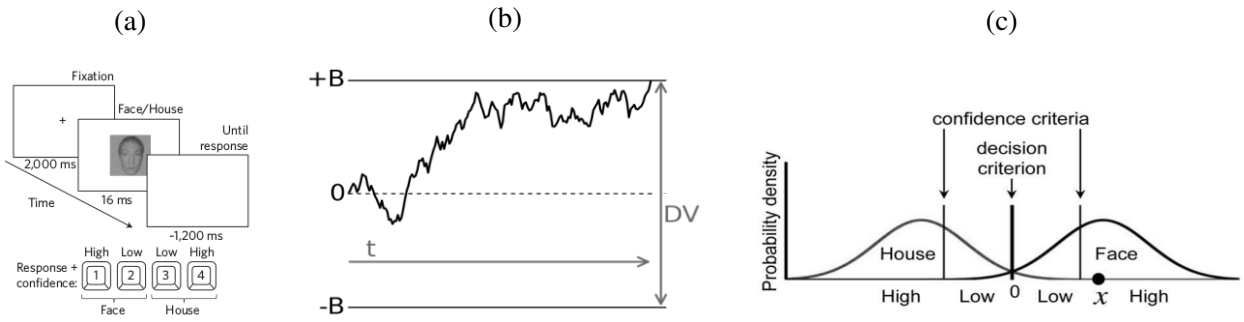


Figure 2.2: **Computational models of perceptual decision making. a)** In Perceptual decision making studies, the subject should infer the hidden state of the environment based on noisy sensory data presented for a very short time. The decisions are usually binary. In some studies, subjects should also report their confidence about their choice. **b)** In the Drift-Diffusion Model (DDM), the decision variable (DV) is the sum of observations over time. The process stops when the DV reaches one of the decision bounds. **c)** Signal Detection Theory (SDT) is a Bayesian approach that turns both choice and confidence report into comparison to a criterion. For example, if input x observed, the subjects chooses *face* with *high confidence*. (a) and (c) are from [114].

(i.e. the experiment is a reaction time task), or when the subject terminates information gathering early in a trial because the extra information is not worth the cognitive cost of observation gathering [41]. In these situations, in contrast to SDT's fundamental assumption, the distribution of inputs would not be Gaussian any more. In the next chapter we discuss how this violation also affect analyzing confidence with SDT.

2.3 Belief-based decision making in AI

Existence of uncertainty in real world has made MDP impractical in many AI applications such as robotics [150]. As a result, developing an algorithm that can deal with uncertainty is actually one of the important problems in Artificial Intelligence. Partially Observable MDP (POMDP) is

the closest framework to MDP that deals with uncertainty in the environment. It is very similar to MDP except for the addition of observation set and observation function. Formally, POMDP is a tuple $(S, A, Z, T, O, R, \gamma)$ where similar to MDP S is the set of states, A is the set of actions, and $T : |S| \times |A| \times |S| \rightarrow [0, 1]$ is the transition function determining $T(s, a, s') = P(s'|a, s)$. Z is the set of observations and $O : |S| \times |A| \times |Z| \rightarrow [0, 1]$ is the observation function determining probability of observation z after performing action a and ending up in state s', i.e. $O(s', a, z) = P(z|a, s')$. Reward function $R : |S| \times |A| \times |O| \rightarrow \mathbb{R}$ is a bounded function determining the gained reward if action a is performed in state s and z is observed. Finally, γ is the discount factor similar to MDP. Starting from a prior probability distribution over states of the environment, called the initial belief (b_0), the goal is to maximize the expected discounted reward [128]. For POMDP, the optimal decision policy π^* can be expressed as a mapping from belief states (probability distributions over states) to distribution of actions that maximizes the total expected reward [142]:

$$\pi^* = \arg\max_{\pi} \sum_{t=0}^{H} \gamma^t E[R(s_t, a_t, z_{t+1})|b_0, \pi]. \tag{2.6}$$

The uncertainty about the state, makes the agent navigate in the belief state space instead of the state space. At time step t, the belief state b_t is updated based on the previous belief state b_{t-1} after action a_{t-1} and observation z_t as follows:

$$b_t(s) \propto P(z_t|s, a_{t-1}) \sum_{s' \in S} P(s|s', a_{t-1}) b_{t-1}(s') \tag{2.7}$$

The uncertainty about the state also makes the problem of finding the optimal policy exponentially more complex than the MDP. While the optimal policy of an MDP can be found in polynomial time, finding the optimal policy of a POMDP is NP-hard [150]. As a result, the optimal policy could only be approximated by methods such as heuristics [81], sampling [115, 141], and search trees [128]. Also, if the belief state can always be expressed as a distribution with a few parameters, e.g. a Gaussian distribution with 2 parameters (mean and variance), the solution can be approximated by discretization of parameters space and solving an MDP in that space [150]. This approach is especially useful in designing or modeling a task with a POMDP.

2.4 Belief-based decision making framework for psychology and neuroscience

Since animals are quite successful in surviving and decision making in the presence of huge uncertainty in the real world, we suspect that belief-based RL and consequently POMDP framework is an appropriate framework to model their decision making. Applications of POMDP in neuroscience and psychology studies have been very limited [120, 67, 68, 5]. Moreover, the decision maker's belief and internal model has not been investigated thoroughly in these studies. For example, choice and reaction time in perceptual decision making tasks were modeled by the POMDP framework before [68]. These two parameters are also explainable by Drift-Diffusion Models (DDMs) [123]. Our approach, however, predicts phenomena related to the confidence that are not explainable by other existing computational frameworks. In the field of social decision making, POMDP and other probabilistic approaches have explained the human behavior in dyadic interactions [166, 69, 51, 5], but ours is the first to unfold decision making in large groups, mostly due to its focus on the belief and internal model of the decision maker.

While our approach is heavily based on POMDP framework developed in AI, there is one important (mostly conceptual) distinction in the assumptions of POMDP used in AI and ours. In a classic POMDP problem, although the agent does not have access to the current hidden state, it knows the world model perfectly. In other words, S, T, and O are completely aligned with the real world. The brain however, needs to learn the hidden state space, and associated transition and observation function. Therefore, these parameters are from a "learned internal model" and could be different from the real generative model. As a result, some seemingly sub-optimal behaviors could be actually from the difference between learned and real model, not a sub-optimal decision making strategy. Moreover, it is important to consider the plausibility of the learned model given the type of the feedback the decision maker receives (especially in nonverbal animals). Overall, we need to find a plausible state space and likelihood function for the decision maker that produces approximately the same observation distribution. This is basically very similar to Variational Autoencoders (VAEs) [53]. We talk about this point when we model each task with a POMDP in the next sections.

Another important issue is the computational cost of finding the optimal policy for POMDPs. As we mentioned, this is the main barrier in using this framework in AI. One could imagine it could be even more problematic when we hypothesize that POMDP is implemented in the brain. While, our framework is at the abstract level of computation, providing methods for computationally efficient implementation of it would significantly strengthen our hypothesis. As, we show in the next chapters, this is achievable by maintaining a closed form of belief state and using greedy methods for policy computation.

Chapter 3

MODELING CONFIDENCE IN PERCEPTUAL DECISION MAKING

3.1 Introduction

The brain is faced with the persistent challenge of decision making under uncertainty due to noise in the sensory inputs and perceptual ambiguity. A mechanism for self-assessment of one's decisions is therefore crucial for evaluating the uncertainty in one's decisions. This kind of decision making, called perceptual decision making, and the associated self-assessment, called confidence, have received considerable attention in decision making experiments in recent years [79, 78, 116]. One possible way of estimating the confidence of a decision maker is to assume that it is equal to the accuracy (or performance) on the task. However, the decision maker's belief about the chance of success and accuracy need not be equal because the decision maker may not have access to information that the experimenter has access to [41]. For example, in the well-known task of random dots motion discrimination [138], on each trial, the experimenter knows the difficulty of the task (coherence or motion strength of the dots), but not the decision maker [40, 86]. In this case, when the data is binned based on difficulty of the task, the accuracy is not equal to decision confidence. An alternate way to estimate the subject's confidence is to use auxiliary tasks such as post-decision wagering [113] or asking the decision maker to estimate confidence explicitly [87]. These methods however only provide an indirect window into the subject's confidence and are not always applicable.

Here, we explain how a model of decision making based on Partially Observable Decision Making Processes (POMDPs) [121, 67] can be used to estimate a decision maker's confidence based on experimental data. POMDPs provide a unifying Bayesian framework for modeling several important aspects of perceptual decision making including evidence accumulation via Bayesian updates,

---

This chapter is mainly outsourced from [80, 84].

the role of priors, costs and rewards of actions, etc. One of the advantages of the POMDP over the other models is that it can incorporate various types of uncertainty in computing the optimal decision making strategy. Drift-diffusion and race models are able to handle uncertainty in probability updates [40] but not the costs and rewards of actions. Furthermore, these models originated as descriptive models of observed data, where as the POMDP approach is fundamentally normative, prescribing the optimal policy for any task requiring decision making under uncertainty. In addition, the POMDP model can capture the temporal dynamics of a task. Time has been shown to play a crucial role in decision making, especially in decision confidence [87, 63]. POMDPs have previously been used to model evidence accumulation and understand the role of priors [121, 67, 68]. To our knowledge, this is the first time that it is being applied to model confidence and explain experimental data on confidence-based decision making tasks.

In the following sections, we introduce some general concepts in perceptual decision making and show the relationship between confidence and accuracy in an optimal Bayesian decision maker who seeks to maximize the total reward. We then model a well-known experiment in perceptual decision making involving confidence, i.e. a fixed duration motion discrimination task with post-decision wagering [86] with POMDP. In this modeling, we go beyond theory and make realistic assumptions about the internal model of the decision maker.

We conclude by showing that the Bayesian inference component of our POMDP model can be implemented by the neural mechanisms that integrate evidence toward a decision bound, consistent with drift diffusion models (DDMs) [123] or more generally, models based on bounded-accumulation of evidence. The POMDP model commits to a choice when the value of the expected improvement of accuracy with new observations is less than the cost of making those observations. We show that this termination criterion uniquely maps to a time-varying decision bound for integration of evidence in the DDM (see also [68]). Such time-varying bounds match past behavioral studies [122, 124] and can be implemented by the urgency signals observed in electrophysiological recordings [22, 118, 24]. Overall, the neural implementation of inference and choice in our POMDP framework is both simple and plausible.

3.2 Accuracy, belief and confidence in perceptual decision making

Consider perceptual decision making tasks in which the subject has to guess the hidden *state* of the environment correctly to get a *reward*. Any guess other than the correct state usually leads to no reward. The decision maker has been trained on the task, and wants to obtain the maximum possible reward. Since the state is hidden, the decision maker must use one or more *observations* to estimate the state. For example, the state could be one of two biased coins, one biased toward heads and the other toward tails. On each trial, the experimenter picks one of these coins randomly and flips it. The decision maker only sees the result, heads or tails, and must guess which coin has been picked. If they guess correctly, they get a reward immediately. If they fail, they get nothing. In this context, *Accuracy* is defined as the number of correct guesses divided by the total number of trials. In a single trial, if A represents the action (or choice) of the decision maker, and S and Z denote the state and observation respectively, then Accuracy for the choice s with observation z is the probability $P(A = a_s|S = s, Z = z)$ where a_s represents the action of decision maker, i.e. choosing s, and s is the true state. This Accuracy can be measured by the experimenter. However, from the decision maker's perspective, their chance of success in a trial is given by the probability of s being the correct state, given observation z: $P(S = s|Z = z)$. We call this probability the decision maker's $belief$. After choosing an action, for example a_s, the $confidence$ for this choice is the probability: $P(S = s|A = a_s, Z = z)$. According to Bayes theorem:

$$P(A|S,Z)P(S|Z) = P(S|A,Z)P(A|Z). \tag{3.1}$$

As the goal of our decision maker is to maximize their reward, they pick the most probable state. This means that on observing z they pick a_{s^*} where s^* is the most probable state, i.e. $s^* = argmax(P(S = s|Z = z))$. Therefore, $P(A|Z = z)$ is equal to 1 for a_{s^*} and 0 for the rest of the actions. As a result, accuracy is 1 for the most probable state and 0 for the rest. Also $P(S|A,Z)$ is equal to $P(S|Z)$ for the most probable state. This means that, given observation z, accuracy is equal to the confidence on the most probable state. Also, this confidence is equal to the belief of the most probable state. As confidence cannot be defined on actions not performed, one could

consider confidence on the most probable state only, implying that accuracy, confidence, and belief are all equal given observation z:

$$\sum_s P(A = a_s|S = s, Z)P(S = s|Z) = P(S = s^*|A = a_{s^*}, Z) = P(S = s^*|Z).[1] \tag{3.2}$$

All of the above equalities, however, depend on the ability of the decision maker to compute $P(S|Z)$. According to Bayes' theorem $P(S|Z) = P(Z|S)P(S)/P(Z) \quad (P(Z) \neq 0)$. If the decision maker has the perfect observation model $P(Z|S)$, they could compute $P(S|Z)$ by estimating $P(S)$ and $P(Z)$ beforehand by counting the total number of occurrences of each $state$ without considering any observations, and the total number of occurrences of observation z, respectively. Therefore, accuracy and confidence are equal if the decision maker has the perfect model of the observation. Sometimes, however, the decision maker does not even have access to Z. For example, in the motion discrimination task, if the data is binned based on difficulty (i.e., motion strength), the decision maker cannot estimate $P(S|difficulty)$ because they do not know the difficulty of each trial. As a result, accuracy and confidence are not equal.

In the general case, the decision maker can utilize multiple observations over time, and perform an action on each time step. For example, in the coin toss problem, the decision maker could request a flip multiple times to gather more information. If they request a flip two times and then guesses the state to be the coin biased toward heads, their actions would be *Sample, Sample, Choose heads*. they also have two observations (likely to be two $Heads$). In the general case, the state of the environment can also change after each action.[2] In this case, the relationship between accuracy and the confidence at time t after a sequence (history H_t) of actions and observations $h_t = a_0, z_1, a_2, ..., z_{t-1}, a_{t-1}$, is:

$$P(A_t|S_t, H_t)P(S_t|H_t) = P(S_t|A_t, H_t)P(A_t|H_t). \tag{3.3}$$

With the same reasoning as above, accuracy and confidence are equal if and only if the decision maker has access to all the observations and has the true model of the task.

[1] In the case that there are multiple states with maximum probability, Accuracy is the sum of the confidence values on those states.

[2] In traditional perceptual decision making tasks such as the random dots task, the state does not usually change. However, our model is equally applicable to this situation.

3.3 Modeling perceptual decision making with POMDPs

Results from experiments and theoretical models indicate that in many perceptual decision making tasks, if the previous task state is revealed, the history beyond this state does not exert a noticeable influence on decisions [41], suggesting that the Markov assumption and the notion of belief state is applicable to perceptual decision making. Additionally, since the POMDP model aims to maximize the expected reward, the problem of guessing the correct state in perceptual decision making can be converted to a reward maximization problem by simply setting the reward for the correct guess to 1 and the reward for all other actions to 0. The POMDP model also allows other costs in decision making to be taken into account, e.g., the cost of sampling, that the brain may utilize for metabolic or other evolutionary-driven reasons. Finally, as there is only one correct hidden state in each trial, the policy is deterministic (choosing the most probable state), consistent with the POMDP model. All these facts mean that we could model the perceptual decision making with POMDP framework. In the cases where all observations and the true environment model are available to the decision maker, the belief state in the POMDP is equal to both accuracy and confidence as discussed above. When some information is hidden from the decision maker, one can use a POMDP with that information to model accuracy and another POMDP without that information to model the confidence. In the next section, we investigate the applicability of the POMDP model in the context of a well-known task in perceptual decision making considering these mentioned points.

3.4 Motion direction discrimination with post-decision wagering experiment

We tested our framework using behavioral data from two monkeys performing a direction discrimination task with post-decision wagering (Fig. 3.1) [86]. On each trial, monkeys observed a patch of randomly moving dots [15] and decided about the net direction of motion. The difficulty of the decision was varied randomly from trial to trial by changing the percentage of coherently moving dots (the "motion strength" or "coherence") and the duration of the motion stimulus (Fig. 3.2a). The stimulus was followed by a delay period and at the end of the delay, the fixation point dis-

appeared (Go cue), signaling the monkey to report its choice with a saccadic eye movement. On a random half of trials, the monkey was given only the right and left direction targets. Choosing the correct motion direction (right target for rightward motion and left target for leftward motion) resulted in a large reward (a large drop of juice) but choosing the incorrect target resulted in no reward and a short timeout. On the other half of trials, the monkey was offered a third target, in addition to the direction targets, in the middle of the delay period. This third target was a sure-bet option. The monkey could choose either the direction targets or the sure-bet after the Go cue. Choosing the sure-bet target guaranteed reward but the magnitude of the reward (volume of the juice) was smaller than that for choosing the correct direction target.

An optimal decision maker who desires to earn more reward and maximize utility should choose the risky, high-paying direction targets when confident about motion direction and the sure-bet option when doubtful about the correct direction. Monkeys showed a similar behavioral pattern. They chose the sure-bet option more often on more difficult trials, where motion strength was low or motion duration was short (Fig. 3.2b). Further, when they ignored the sure-bet option and chose the high-stakes direction targets, their accuracy was higher compared to the trials with similar difficulty without the sure-bet option when they had to choose one of the direction targets (trials without sure-bet target; Fig. 3.2a). These results indicate the presence of a mechanism for assessment of expected decision outcome (confidence), and reliance on this mechanism for guiding the opt-out behavior.

3.5 POMDP model of the direction discrimination task

The motion direction discrimination task has previously been modeled using the POMDP framework [121, 41, 68]. However, in these models, the subject's confidence was either not modeled[121, 68] or was obtained assuming the subject had an exact generative model of the task [41]. Such knowledge, however, is unlikely in most natural contexts and common task designs. For example, in the direction discrimination task, subjects face a mixture of stimulus difficulties across trials. They neither know the exact generative function for the stimulus on each trial nor the exact set of motion strengths used in the experiment.

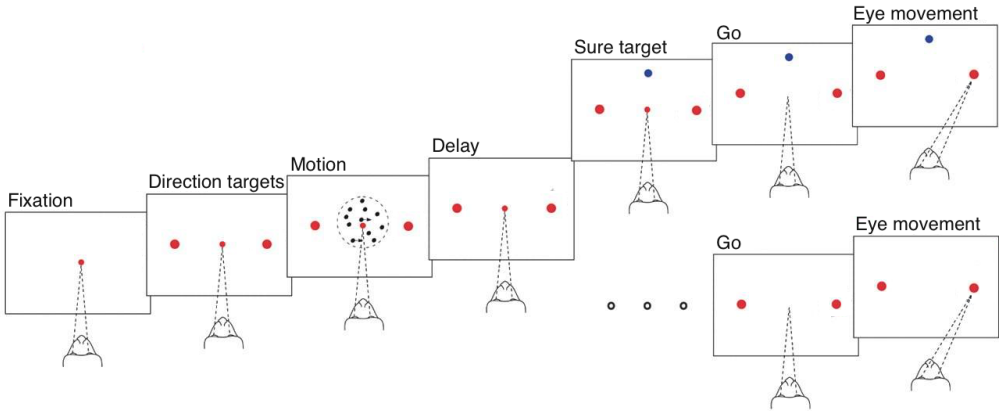


Figure 3.1: **Motion direction discrimination with post-decision wagering task design.** On each trial, monkeys viewed a patch of randomly moving dots and made a decision about the net direction of motion. Stimulus strength and duration varied randomly from trial to trial. On half of the trials, only the right and left direction targets were shown (large red dots). The motion stimulus was followed by a delay period. The central fixation point (small red dot) disappeared at the end of the delay (Go cue), instructing the monkey to report perceived motion direction with a saccadic eye movement to one of the two direction targets. Choosing the correct target (right for rightward motion and left for leftward motion) yielded a large reward, whereas choosing the incorrect target resulted in a short timeout. On the other half of the trials, a third target (sure target, shown as a blue dot) appeared on the screen during the delay period. Choosing this target after the Go cue yielded a guaranteed but smaller reward than choosing the correct direction target.

Following previous models [121, 41], we define the hidden state of the environment for our POMDP to include both the unknown direction and unknown coherence, combined into a single real-value which we call "signed motion coherence" c: positive values of the signed motion coherence indicate rightward motion and negative values indicate leftward motion [132]. Specifically, the momentary observations z_t at times t for a trial with signed coherence c are modeled as samples independently drawn from a Gaussian distribution, $\mathcal{N}(c, w_z)$, with mean $\mu = c$ and variance w_z^2 (Fig. 3.3a).

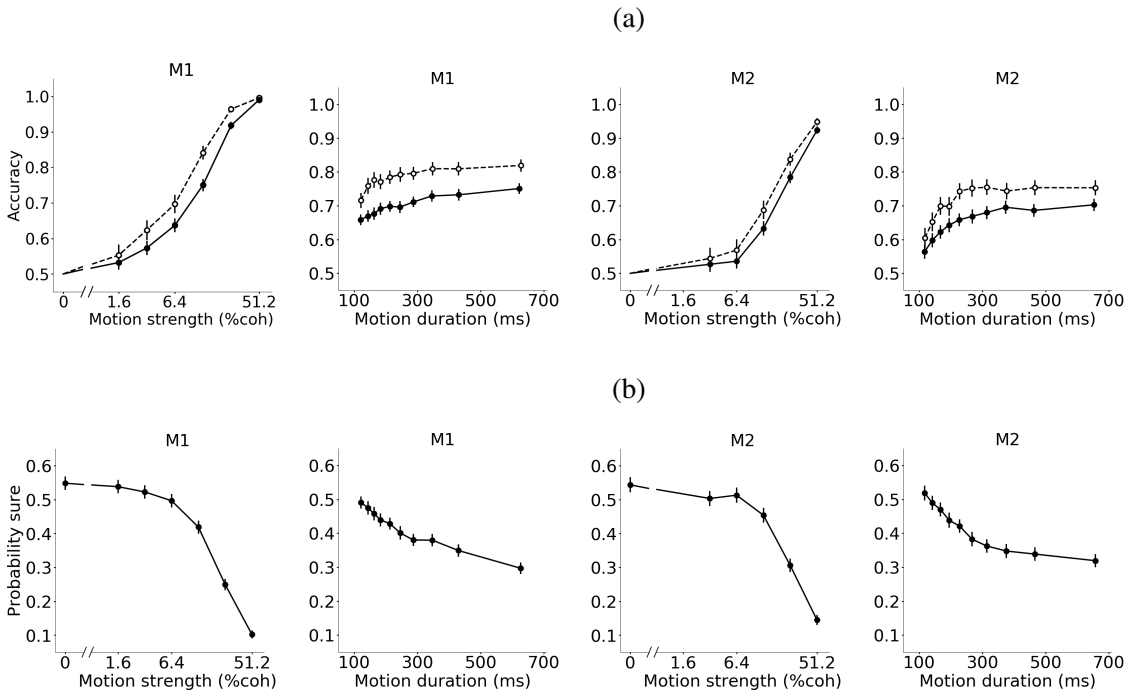


Figure 3.2: **Monkeys' behavior in random dots task with post-decision wagering.a**) Accuracy as a function of motion strength and duration for the two monkeys (M1 and M2). Solid lines show the accuracy on trials where the sure target was not presented. Dashed lines show the accuracy on trials where the sure target was shown but the monkey chose one of the high-stakes direction targets. **b)** Probability of choosing the sure target for different motion strengths and durations for monkeys M1 and M2. Error bars indicate standard error of the mean (s.e.m.).

The two main actions of our POMDP model are committing to direction *right* or direction *left*. Also, action "observe" makes the next observation available to update the model's belief about c. Finally, the action of choosing the sure-bet option is available during the delay period on half of the trials. The decision maker gets r_{right} as the reward utility for committing to direction *right* if and only if the direction of the hidden state is *right* ($c > 0$). r_{left} is the reward utility given to the decision maker by committing to direction *left* if and only if the direction of the hidden state is *left* ($c < 0$). Choosing the sure-bet option, if available, always yields reward utility of r_{sure}.

The POMDP model begins each trial with a prior belief about the signed coherence of the trial. Subjects are not explicitly informed about the exact set of discrete motion coherence levels used in the experiment. They only experience largely overlapping distributions of motion energies on different trials [85]. Therefore, it is most realistic to consider that the model's prior spans a continuous domain, obtained from observations across all trials with various coherence levels and durations. Because the logarithmic spacing of the discrete motion coherences used in the experiments $(0\%, 1.6\%, 3.2\%, 6.4\%, 12.8\%, 25.6\%, 51.2\%)$ causes the mass of the prior distribution to be largely concentrated in its central peak around 0, our POMDP model uses a Gaussian approximation to this prior distribution, $\mathcal{N}(0, \sigma_0)$ (Fig. 3.3b).

Starting with a Gaussian prior (initial belief) $b_0 = \mathcal{N}(\mu_0 = 0, \sigma_0)$, the model iteratively updates its belief about the hidden state of the environment, i.e., the signed motion coherence c, following each observation, z_t, drawn from the distribution $\mathcal{N}(c, w_z)$ at time step t (Fig. 3.3c). To be able to update the belief, knowledge of the true observation variance, w_z^2, is required. However, w_z^2 is unknown to the model. Rather, we use σ_z^2 to denote the model's learned observation variance. This means that the model assumes z_t is drawn from the Gaussian likelihood function $P(z_t|c) = \mathcal{N}(z_t; c, \sigma_z)$. A Gaussian prior and a Gaussian likelihood function together result in a Gaussian posterior [101, 41] (Fig. 3.3c) for c given by:

$$
\begin{aligned}
&b_t = P(c|z_1, \ldots, z_t) = \mathcal{N}(\mu_t, \sigma_t) \\
&\mu_t = \frac{\sigma_{t-1}^2 z_t + \sigma_z^2 \mu_{t-1}}{\sigma_{t-1}^2 + \sigma_z^2} = \frac{\sigma_z^{-2}}{t\sigma_z^{-2} + \sigma_0^{-2}} \sum_{j=1}^{t} z_j \qquad \sigma_t^2 = \frac{\sigma_{t-1}^2 \sigma_z^2}{\sigma_{t-1}^2 + \sigma_z^2} = \frac{1}{t\sigma_z^{-2} + \sigma_0^{-2}}
\end{aligned}
\tag{3.4}
$$

Since the reward only depends on choosing the correct motion direction, the POMDP model's choice depends on μ_t, and consequently $\sum_{j=1}^{t} z_j$, being larger than zero for choosing the rightward direction and less than zero for choosing the leftward direction. A random choice is made in the unlikely event that μ_t is exactly equal to 0. Moreover, the model's confidence is the posterior probability of the chosen direction, which is the sum of the posterior probabilities over all motion coherences in that direction, i.e., $\Phi(\mu_t/\sigma_t)$ when $\mu_t \geq 0$ and $\Phi(-\mu_t/\sigma_t)$ when $\mu_t < 0$, where $\Phi(x)$ denotes the standard normal cumulative distribution function [41].

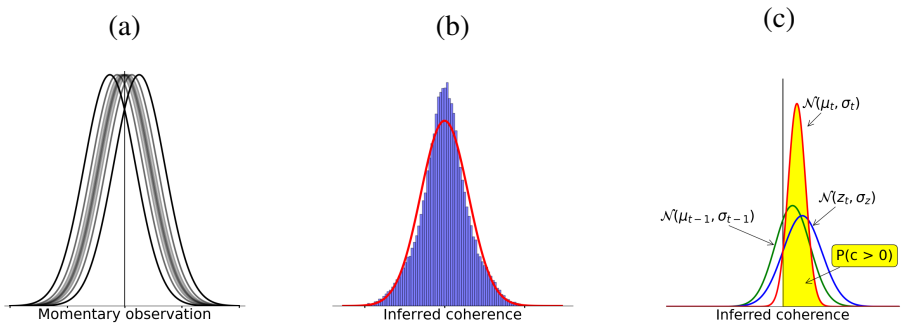


Figure 3.3: **The POMDP model of the direction discrimination task. a)** Probability distribution of momentary observations for a motion coherence c is modeled as a Gaussian distribution with mean $\mu = c$ and variance w_z^2. There are multiple Gaussian distributions for different motion coherence. Positive and negative observations indicate rightward and leftward motion directions, respectively. **b)** The distribution of inferred coherence across all trials provides the initial belief state of the POMDP model (blue histogram) at the beginning of each trial. The initial belief is approximated by a Gaussian function (red curve). **c)** The POMDP model sequentially updates its belief about the motion coherence based on new observations. Combining the belief at time $t-1$ (green distribution) with the acquired observation from the stimulus at time t (blue distribution) results in a new belief at t (red distribution). The expected likelihood that the rightward choice is correct is the area under the updated belief distribution for positive sensory evidence (yellow region).

The POMDP approach can easily model termination of the decision-making process and commitment to a choice by assigning a cost (negative utility) to observation gathering and belief update (via the action "observe")[121]. Moreover, because the hidden state does not change with actions within a trial in the motion discrimination task, a one-step look-ahead search [128] is adequate to determine the optimal decision policy for non-decreasing observation costs over time (instead of computing the total expected reward utility to the end of the trial; see the proof in 3.5.1). The

model halts new observations when the expected increase in confidence is less than the ratio of the cost of an observation and the reward utility for correct choice. The expected increase in confidence after one more observation depends on the current belief and the probability distribution of the next observation according to the model. Specifically, when the current belief is $\mathcal{N}(\mu_t, \sigma_t)$, the model assumes that the next observation is a sample from $\mathcal{N}(\mu_t, \sigma_z)$, where σ_z^2 is the learned observation variance. Figure 3.4a shows the expected increase in confidence for a new observation as a function of two key variables: the inferred μ_t and the elapsed time. The expected increase in confidence from new observations is higher earlier in the trial and for smaller inferred mean coherence, μ_t.

A constant observation cost over time, if present, would give rise to a stopping criterion that matches an iso-gain contour. These contours would effectively implement a time-varying bound on μ_t for each motion direction (an upper bound and a lower bound). Figure 3.4b shows these collapsing bounds for a cost of 10^{-3} per observation (in our case, per 10 ms) when the reward utility for a correct direction choice is set to 1. A policy for termination of observations is especially critical in reaction time (RT) tasks where subjects have to decide when to initiate a response. However, a termination policy could exist even in tasks where stimulus duration is controlled by the experimenter, causing early termination of the subject's decision-making process before stimulus offset, especially in long and easy trials [85, 106].

The reward utility maximization principle also determines the choice when the sure-bet option is available. As the reward for the sure-bet option is guaranteed, the POMDP model compares the expected reward utility for choosing each direction with the reward utility for the sure-bet option in order to pick the final action:

$$a_t = \begin{cases} left & b_{t,left} \cdot r_{left} > b_{t,right} \cdot r_{right} \text{ and } b_{t,left} \cdot r_{left} > r_{sure} \\ right & b_{t,right} \cdot r_{right} > b_{t,left} \cdot r_{left} \text{ and } b_{t,right} \cdot r_{right} > r_{sure} \\ sure & r_{sure} \geq b_{t,right} \cdot r_{right} \text{ and } r_{sure} \geq b_{t,left} \cdot r_{left} \end{cases} \tag{3.5}$$

Because $r_{right} = r_{left} = r_{direction}$ in our task, the above policy reduces to a comparison of the model's confidence with the reward utility ratio $r_{sure}/r_{direction}$ between the sure-bet and correct

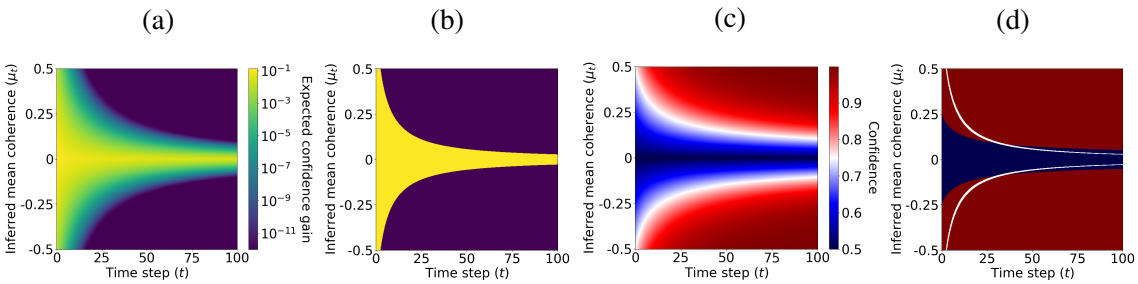


Figure 3.4: **Computation of choice and confidence in the POMDP model. a)** The expected confidence gain for a new observation as a function of inferred mean coherence, μ_t, and elapsed time, t. **b)** An example POMDP decision policy when new observations are associated with a constant cost. The yellow area represents the belief states where the optimal action is to continue observing. The purple area represents the belief states where the POMDP model terminates and commits to a choice. **c)** Confidence as a function of inferred mean coherence, μ_t, and time, t. **d)** The ratio of reward utilities for sure-bet and correct direction choices determines the POMDP policy for choosing the sure-bet option. The policy for sure-bet can be illustrated as phase boundaries in the confidence plot of panel g. The blue region denotes combinations of inferred coherence and time for which the model would choose the sure-bet target. The red region denotes (μ_t, t) for which direction targets are chosen. Thresholds for separating low and high confidence ratings are thus the boundaries between blue (low confidence) and red (high confidence) regions. Solid white lines show the two decision termination bounds where the model stops gathering more observations and commits to a decision. In the simulations, $\sigma_z = 2.0$, $\sigma_0 = 1.0$, and the utility ratio = 0.63.

direction choices. Since confidence increases with the absolute value of inferred coherence, $|\mu_t|$, this reward utility ratio leads to a time-varying boundary that determines the POMDP policy as a function of inferred coherence and time in each direction (upper and lower bounds). Figure 3.4c shows confidence as a function of inferred coherence and elapsed time for an example POMDP model and Figure 3.4d shows the model policy for an example reward utility ratio of 0.63.

With a constant observation cost, the model has up to four degrees of freedom: (*i*) observation cost; (*ii*) the true observation variance (w_z^2), which shapes input samples available to the model; (*iii*) the learned observation variance (σ_z^2), which the model attributes to its inputs; and (*iv*) the learned variance of the prior distribution (σ_0^2). For an optimized POMDP model, however, σ_0^2 and σ_z^2 are uniquely determined by w_z^2 and observation cost. As mentioned before, σ_0^2 determines the prior belief, which should be consistent with the overall distribution of states and consequently, perceived observations. Moreover, σ_z^2 should match the model's posterior belief with its average accuracy for each motion duration. This is possible based on the feedback given about motion direction choices (correct or wrong) after each trial (see the next section for details on estimating these parameters). Such a model, therefore, has two degrees of freedom: observation cost and w_z^2.

Note that correct posterior belief (matched with accuracy on average) is not necessary for maximizing the reward utility in choosing between the two directions because determining the sign of the sum of observations is sufficient. However, it is necessary for the wagering task where the expected reward utility of choices needs to be computed (Eq. 3.5).

3.5.1 One-step look-ahead search as the optimal strategy

Here we show that for an unbiased 2-alternative decision-making task such as ours, one-step look ahead search results in the optimal POMDP policy for a non-decreasing observation cost over time. First, note that due to the symmetry of the task for direction choices, the optimal decision maker picks the choice with the highest belief. This means that when considering whether to terminate or continue acquiring observations, an optimal decision maker compares the observation cost and the resultant expected confidence (belief).

Second, the entropy, i.e. $-b_{right}\log(b_{right}) - b_{left}\log(b_{left})$, has an inverse relationship with confidence. The expected information gain (i.e., decrease in entropy) decreases with more samples (here observations) [58]. As a result, the expected increase in confidence decreases with the number of observations as well. This means that if the expected increase in confidence with one more observation is less than the cost of the observation, the expected increase in confidence with k more observations is less than k times the cost of one observation. Thus, if the cost of observations

is non-decreasing over time, comparing the expected confidence with the cost of an observation at the current time is enough to maximize the expected total reward. In other words, if the next observation is not worth its cost, making more observations would not be worth the cost either. Importantly, this holds for any observation function and state space as long as the probability distribution for observations does not change with time, which is true in our task (coherence does not change within a trial).

3.6 Comparison of model predictions with experimental data

In our task, the stimulus viewing duration was controlled by the experimenter and subjects were required to maintain fixation throughout the duration. As a result, the cost of acquiring new observations while maintaining fixation on the stimulus could be negligible. We verified this hypothesis by comparing the model with two degrees of freedom (observation cost and w_z) to a POMDP that uses all observations in each trial with only w_z as the free parameter). They were not significantly different in quality of fits even without penalizing the extra free parameter (Vuong's closeness test [156], $p = 0.16$ for monkey M1 and $p = 0.07$ for monkey M2).

We fit the model to each monkey's accuracy on trials in which the sure target was not shown (Fig. 3.5a) ($R^2 = .95$ and .88 for monkeys 1 and 2, respectively) and obtained the observation variance w_z^2. Specifically, when there is no observation cost, the average belief about the direction *right* is $\Phi(\sqrt{t}c/w_z)$ for trials with duration t and signed coherence c. Therefore, as we did not have access to observations in each trial, we modelled the probability of choosing the direction *right* with a Bernoulli distribution whose mean is $\Phi(\sqrt{t}c/w_z)$ (when c is negative, the probability of choosing the direction *right* becomes less than .5).

Each monkey's data were fit separately. For monkey M1, w_z was .90 while for monkey M2, it was 1.69. Based on these w_z values, we estimated the prior belief $b_0 = \mathcal{N}(\mu_0, \sigma_0)$ as follows: for any trial with true coherence c and duration t, we generated a sample from $\mathcal{N}\left(c, w_z/\sqrt{t}\right)$; the samples generated from all the trials were used to fit the Gaussian $\mathcal{N}(\mu_0, \sigma_0)$ via maximum likelihood estimation [101].

To calculate σ_z, we fit the POMDP model's confidence $\Phi(|\mu_t|/\sigma_t)$ to the accuracy in all trials

that the sure-bet option was not offered, using w_z and σ_0 estimated as above. In each trial, we calculated $\sum_{i=1}^{t} z_i$, the sum of the observations generated from the actual coherence and the stimulus duration used in that trial. Using the relationship between the sum of observations, μ_t and σ_t in equation 3.4 we get $\Phi(\sigma_z^{-2}\sum_{i=1}^{t} z_i/\sqrt{t\sigma_z^{-2}+\sigma_0^{-2}})$ as the subject's belief about the direction *right*. We calculated a maximum likelihood estimate of σ_z by fitting this belief to the accuracy in all trials where the sure-bet option was not offered. For the fitting, the direction *right* choice was modeled as a Bernoulli distribution whose mean is $\Phi(\sigma_z^{-2}\sum_{i=1}^{t} z_i/\sqrt{t\sigma_z^{-2}+\sigma_0^{-2}})$, where the z_i were sampled based on the true coherence and duration used in the trials.

One can also try to make the fit more accurate by estimating σ_z and σ_0 iteratively. We can start with the values of σ_0 and σ_z obtained as described above, and then readjust σ_0 based on this estimated σ_z. The readjusted σ_0 can be used to fit σ_z again. With every such iteration, we found that the change in σ_0 decreased. We repeated this process until the change in σ_0 became less than our precision error. This process converged in less than 5 iterations for both monkeys. However, the readjusted σ_0 values did not significantly improve the goodness of fit of the belief to the monkey's choice. Nonetheless, we used these more accurate values in our models: σ_0 was .46 and .87, and σ_z was 1.60 and 3.59 for monkey M1 and M2, respectively.

An important observation from our model fitting process is that the estimated σ_z is larger than w_z in both monkeys - this is due to the structure of the task. The task involved a discrete set of seven coherence levels, $\{0\%, 1.6\%, 3.2\%, 6.4\%, 12.8\%, 25.6\%, 51.2\%\}$, and the belief for each direction is the average probability of that direction over these coherence levels [42]. Due to the greater number of low coherence trials compared to high coherence trials, the continuous model with $\sigma_z = w_z$ generates higher confidence compared to a model with the true generative model with seven discrete coherence levels, resulting in an overall overconfidence for each stimulus duration. Therefore, a continuous model that matches its confidence with a subject's accuracy for each stimulus duration considers each observation less reliable, i.e., $\sigma_z > w_z$. A second observation is that although the POMDP policy is deterministic, the stochasticity needed to fit the trial-by-trial choice data comes from the distribution of observations given the true stimulus.

Having estimated the model parameters based on trials without the sure-bet target, we predicted

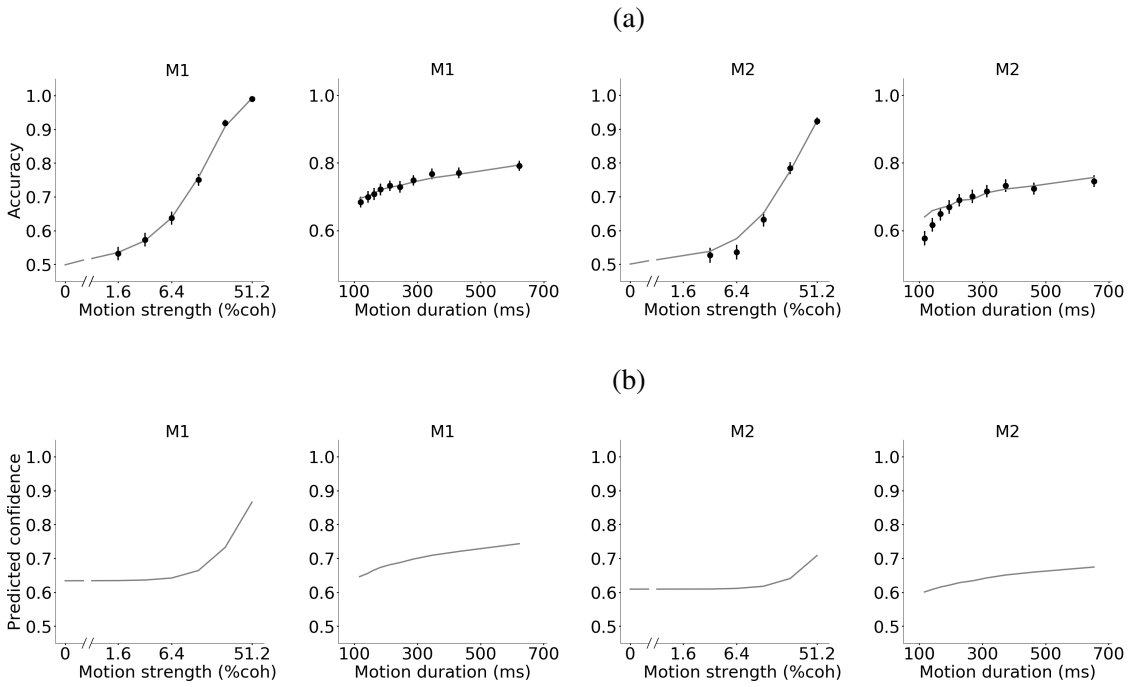


Figure 3.5: **The POMDP model predicts monkeys' confidence from their performance. a)** The model was fit to each monkey's accuracy on trials without the sure-bet option. Solid lines are model fits and data points are the measured accuracy for each motion strength and duration for monkeys M1 and M2. **b)** The model parameters obtained from the fits in (a) were used to predict confidence for each motion strength and duration for each monkey. Error bars indicate s.e.m.

the monkey's confidence for each motion coherence and duration (Fig. 3.5b). These predictions suggested a monotonic increase in confidence with motion coherence and duration, compatible with previous studies [155, 6, 86, 62, 169].

Since the model chooses the sure-bet option when confidence (belief) is less than the reward utility ratio of the sure-bet and correct direction choice (Eq. 3.5), it predicts lower probability of choosing the sure-bet target on trials with stronger motion and longer durations. Since we do not know the exact utility of reward volumes associated with the sure-bet and correct direction choices, we added a new free parameter to the model that represented the reward utility ratio and used this

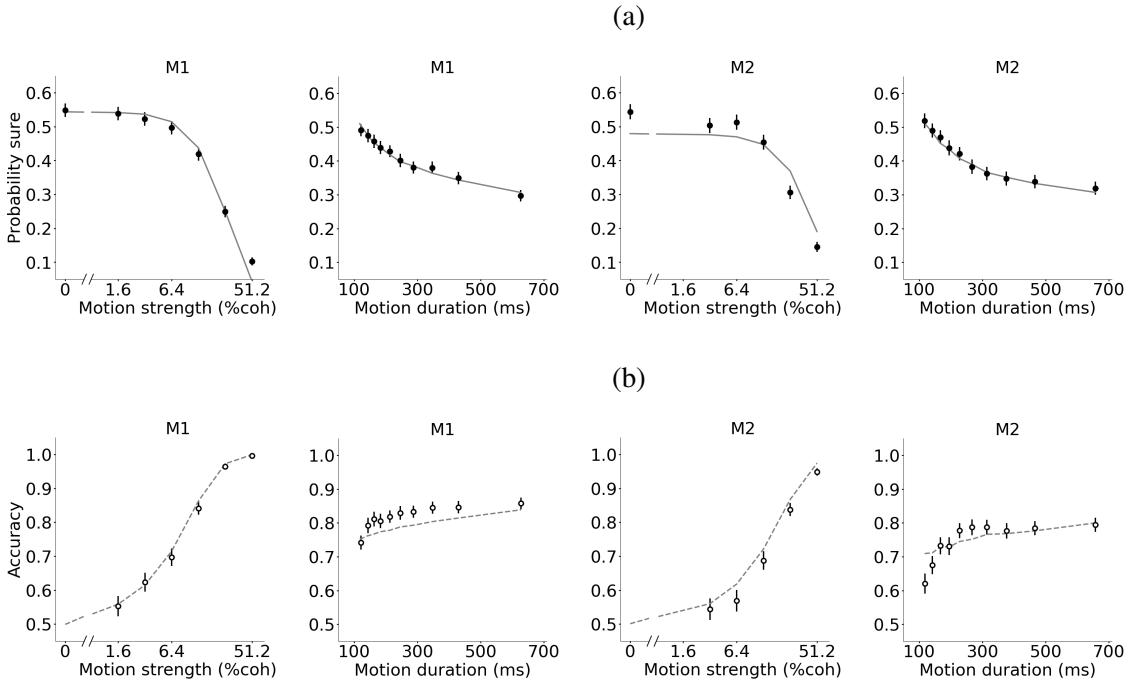


Figure 3.6: **The POMDP model captures the monkeys' behavior based on their confidence.a)** Predictions of the POMDP model about confidence were thresholded to fit the likelihood of choosing the sure-bet option. **b)** With the model parameters fully constrained by accuracy on trials without the sure-bet target and the likelihood of choosing the sure-bet target when it was presented to the monkey, we predicted the monkey's accuracy on trials in which the sure-bet target was shown but ignored. Lines are model predictions. Data points are identical to those in Figure 3.2. Error bars indicate s.e.m.

parameter as a threshold that the confidence was compared to on trials in which the sure-bet target was presented. Optimizing this parameter (.63 for monkey and .59 for monkey 2) in order to match the predicted confidence of the POMDP model with the monkey's behavior provided a fit with $R^2 = 0.90$ and 0.82 for monkey M1 and monkey M2, respectively (Fig. 3.6a).

Since the model parameters are fully specified based on the monkey's accuracy on trials without the sure-bet target and the probability of choosing the sure-bet target when it was presented, we could provide quantitative predictions for the monkey's direction choice accuracy when the sure-bet target was presented but not chosen. Figure 3.6b shows these predictions (gray dashed lines), demonstrating that they closely match experimentally measured accuracy on trials where the monkey ignored the sure-bet option ($R^2 = 0.90$ and 0.81 for monkey M1 and monkey M2, respectively). Trials with 0% coherence were removed from this accuracy analysis because a correct direction choice is undefined on those trials and the monkey was rewarded randomly.

3.6.1 Trail-by-trial model fitting compared to batch fitting

Our results are based on fitting our POMDP model parameters to data from all the trials taken together ("batch estimation") as opposed to estimating the parameters iteratively in a trial-by-trial manner. This raises the following concern: does a trial-by-trial fitting process lead to a different estimate for σ_z and change the results?

To address this concern, we estimated model parameters iteratively for both monkeys M1 and M2 (see Figure 3.7a). Parameters were obtained based on maximum likelihood estimation with gradient descent. In the batch approach described in the main text, the gradient was calculated for all trials, and the optimization was performed in a few iterations. In the trial-by-trial approach, the gradient was calculated only for the current trial, which made the optimization process akin to stochastic gradient descent (SGD). To be more consistent with the subject's experience, each trial was used only for one iteration. We found that the results from SGD (or indeed gradient descent based on other unbiased sampling of the data set, e.g., mini batches) converged to the results from applying gradient descent on the entire dataset (Figure 3.7a). This is expected because of the large number of trials. Similar to the batch approach, SGD also produced $\sigma_z > w_z$, even when the

initial values of σ_z and w_z were equal. Thus, an iterative approach leads to essentially the same conclusions as the batch approach.

Another possible concern is that we know neither the subject's true update strategy nor the true observations in each trial used for updating σ_z. However, it is reasonable to assume that similar to SGD, the subject increases σ_z (decreases overall confidence) when it receives negative feedback and decreases σ_z (increases overall confidence) when it receives positive feedback. The amount of update would depend on the actual observations in that trial (e.g., greater increase in σ_z when observations strongly indicating a particular choice result in negative feedback). Thus, although any trial-by-trial analysis is inherently limited by our lack of knowledge about the animal's true observations and update rule, we believe our SGD-based procedure is a reasonable approximation.

Finally, the data we analyzed were from the stable phase of data collection following extensive training with tens of thousands of trials with the same task structure. We therefore tested the effect of trial-by-trial estimation of model parameters after setting the initial value of parameters equal to the values from the batch approach, assuming that the subjects learned them during training. Figure 3.7b shows the evolution of σ_z and w_z with these more reasonable initial values, demonstrating the stability of these parameters across all trials in the experiment. Moreover, the evolution of overall accuracy in Figure 3.7c shows that this was not an artifact of the fitting process. As shown in the figure, σ_z and w_z fluctuated around the initial/batch values and were always within a 0.05 difference range. Accuracy was also quite stable (within a 1% difference range for monkey 1 and 3% for monkey 2).

3.7 Relationship between POMDP and drift diffusion models

A simple mathematical model that has been extensively used to provide quantitative fits to behavior and explain neural activity in various brain regions is the drift diffusion model (DDM) [123]. DDM assumes that each observation confers evidence in favor of one choice and an equal amount of evidence against the other choice (Fig. 3.8a). Integration of sensory evidence over time provides a decision variable (DV) that tracks the total evidence in favor of each choice. In most formulations of DDM, two bounds above and below the initial value of the DV (+B and -B in Fig. 3.8a) act

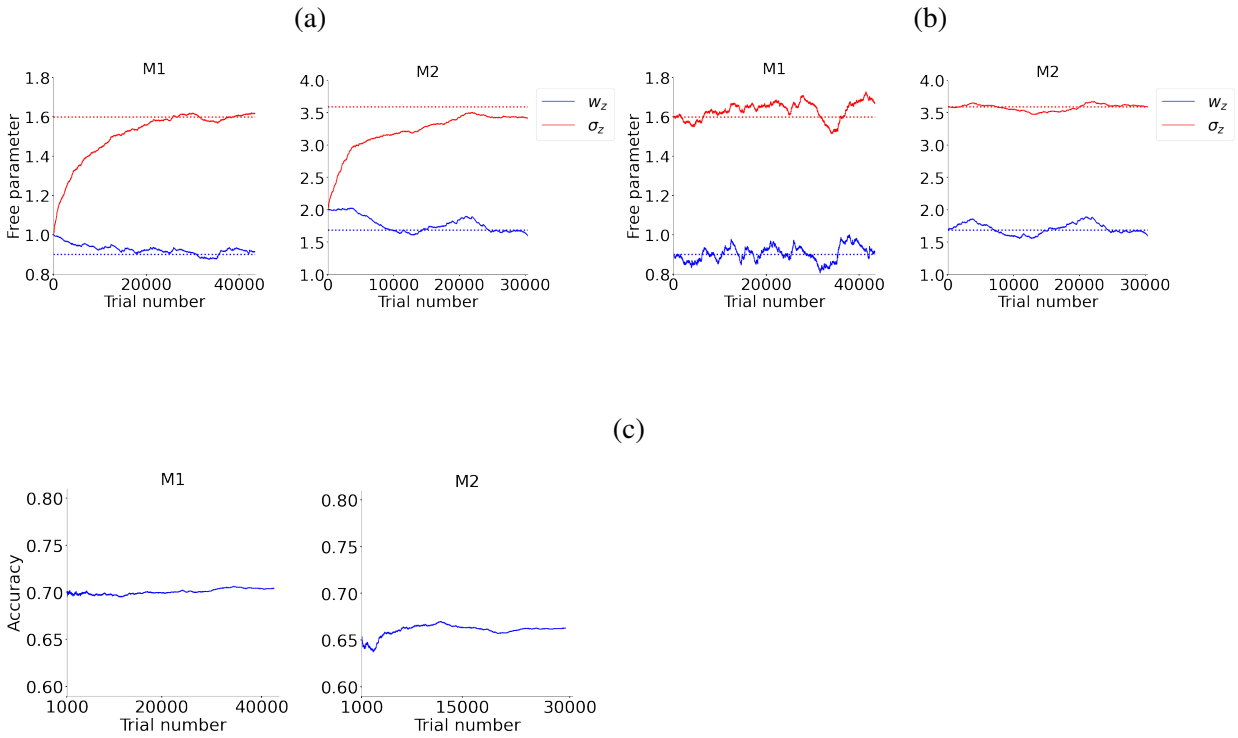


Figure 3.7: **Trial-by-trial model fitting results in similar parameter values as the batch approach.** **a)** Because there are thousands of data points, parameters estimated from trial-by-trial (online) model fitting (red and blue lines) converge to similar parameters as the batch approach (dotted lines) at the end of the trials. **b)** Since our analyzed data is from the stable phase of data collection where the monkeys have already learned the task, it is more realistic to assume that the initial parameters are close to the parameters from the batch approach. For this more realistic initialization, the parameters remain within close range of the initial values throughout the entire experiment. **c)** Stability of overall accuracy indicates that this is not an artifact of our fitting process.

as termination criteria for the decision. As soon as the DV reaches one of these bounds, the decision-making process stops and the choice associated with the bound is made. In cases where the stimulus terminates before a bound is reached, the choice with the most supporting evidence is

selected [85]. For the direction discrimination task, the decision variable, V_t, is updated with each new sensory observation according to:

$$V_t = V_{t-1} + z_t \tag{3.6}$$

where V_{t-1} is the DV at time $t-1$ and z_t represents the momentary sensory observation drawn from a Gaussian distribution with mean c and variance w_z^2. V_0 is initialized to zero when the prior probability and expected reward of the two choices are equal. Therefore, prior to reaching a bound, V_t equals the sum of observations $\sum_{j=1}^{t} z_j$ at time t.

The DDM as described above has previously been linked to probabilistic reasoning between two categories as in signal detection theory [56, 34]. These previous models explain choice accuracy but the subject's belief when there are multiple motion coherence levels was not addressed. A later model by Drugowitsch et al.[41] addressed this issue by adding Bayesian reasoning on the drift rate of the DDM but the generative model was assumed to be known and exact.

Our POMDP model allows for both a learned generative model and a belief update rule that can be mapped to the DDM. Taking the ratio of the two update rules in Equation 3.4, we obtain:

$$\frac{\mu_t}{\sigma_t^2} = \sigma_z^{-2} \sum_{j=1}^{t} z_j = \sigma_z^{-2} V_t \tag{3.7}$$

where the second equality is based on the definition of the DV in DDM (Eq. 3.6). Thus, the Bayes update of the inferred coherence, μ_t, can be achieved via addition in the DDM. This means that there is a unique mapping from μ_t and σ_t^2 of a POMDP model to the V_t and t of the DDM.

This mapping holds in the presence of a bound in the DDM [98, 42]. Moreover, the termination criterion in the POMDP model translates to a unique bound in the DDM. As shown in Figures 3.4b and 4.3a, the policy in the POMDP model can be expressed as a bound in the space defined by inferred mean coherence, μ_t, over time. This bound on μ_t has an equivalent bound on V_t in the DDM. In general, if $\Theta'(t)$ is the time-varying termination criterion applied to μ_t in a POMDP model (as in Fig. 3.4b), the equivalent bound, $\Theta(t)$, on V_t in the DDM is given by:

$$\Theta(t) = \frac{\Theta'(t)}{\sigma_z^{-2}\sigma_t^2} = \left(t + \frac{\sigma_z^2}{\sigma_0^2}\right)\Theta'(t) \tag{3.8}$$

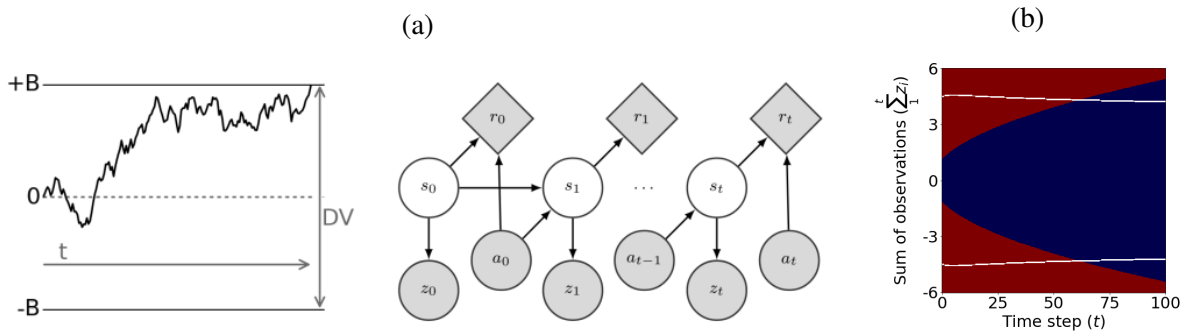


Figure 3.8: **The POMDP policy can be implemented by a DDM with collapsing bounds. a)** (Left panel) In the standard DDM, the decision variable (DV) is the sum of observations over time. The process stops when the DV reaches one of the static decision bounds (+B or -B). (Right panel) Graphical model for a POMDP. For each time step (indicated by the subscripts 0, 1, ..., $t-1$, t), r is the reward gained due to action a in hidden state s. z is the observation in hidden state s. The POMDP model infers a posterior probability distribution over hidden states at each time step based on past observations and actions. In the motion discrimination task, the actions are committing to a choice or making another observation. The model commits to a choice when the expected increase in the probability of a correct response is not worth the cost of an extra observation. **b)** The time-varying bounds on μ_t in the POMDP policy map (e.g., solid white lines in Fig. 4.3a) have equivalent time-varying bounds on the DV in the DDM (Eq. 3.8; white lines in this panel). Similarly, the low and high confidence regions (blue and red regions respectively) of the POMDP policy map in Fig. 4.3a have equivalents in the DDM as shown here.

where the first equality derives from Equation 3.7 and the second from Equation 3.4. Similarly, confidence ratings can be expressed as time-varying boundaries in the DDM. Figure 3.8b shows the decision bound and confidence rating boundaries based on the accumulated evidence in the DDM derived to match the POMDP model in Figure 4.3a.

Overall, both the inference process and the termination criterion of the POMDP model can be

implemented with a DDM, suggesting that the neural circuitry for integration of sensory evidence could effectively be implementing the POMDP policy explained in this chapter.

3.8 Discussion

We present a Bayesian framework based on partially observable Markov decision processes (POMDPs) that accounts for choice and confidence in perceptual decision-making tasks. Our framework explains the effects of observation cost and task structure on choice and confidence. It also elucidates how the observation noise learned by a Bayesian decision maker may systematically differ from the veridical observation noise, and how this difference influences prior beliefs and confidence. We tested our model using the behavioral data of monkeys performing a direction discrimination task with post-decision wagering [86]. The monkeys' choice accuracy provided quantitative predictions about subjective confidence. These predictions fit the monkey's opt-out behavior in our task, indicating that the monkey's confidence matches the POMDP framework. Prediction of confidence purely based on choice accuracy is a remarkable feat, especially considering systematic discrepancies between the two [116]. Further, we show how our POMDP model can be mapped to bounded evidence accumulation models [123] and potentially implemented by the same cortical and sub-cortical neural networks implicated in the decision-making process [57, 135].

We applied the POMDP framework to a fixed-duration task where the stimulus duration was controlled by the experimenter. The framework can also be used to model animal and human behavior in reaction-time tasks [121]. In fact, reaction time tasks might offer better opportunities to study choice and confidence. In fixed-duration tasks, long stimulus durations are least desirable because a multitude of mechanisms, including decision bounds, time-varying attention, or task engagement, could cause partial use of sensory information unbeknownst to the experimenter. Short stimulus durations are not immune to misinterpretations either. Short stimuli can cause neural responses that last longer than the stimulus duration [89, 88, 167], providing an opportunity for selective sampling. Moreover, short-term mnemonic mechanisms and active revision of choice and confidence provide additional opportunities for dissociating the observations used by the decision-making process from those assumed by the experimenter. Reaction-time task designs

where subjects control the stimulus viewing duration, combined with monitoring and manipulation of neural responses in sensory and decision-making circuits, would improve experimental control and enable more accurate interpretation of experimental results [159].

We conclude by noting that simple bounds on decision variables, as employed in traditional models of decision making, might not be sufficient to capture the types of complex policies (mappings of beliefs to actions) required in dynamic environments and in tasks more complex than the random dots task. In such cases, the POMDP model offers a powerful and flexible framework for decision making as it allows (i) arbitrary probability distributions for the prior and observation functions, (ii) arbitrary state transition functions conditioned on the decision maker's actions, and (iii) policies that are not restricted to bounds on decision variables and that implement arbitrary mappings of beliefs to actions [121, 67]. Testing these more general attributes of the POMDP model in animal and human experiments remains an important direction for future research.

Chapter 4

DIVERGENCE OF CONFIDENCE AND ACCURACY IN PERCEPTUAL DECISION MAKING

4.1 Introduction

Confidence plays a key role in guiding behavior in complex environments [34, 86, 117, 155] and is often critical for modeling behavior and understanding its neural mechanisms in such environments [10, 42, 46, 76, 118, 137]. However, unlike sensory choices and their accuracy that are usually easy to measure, confidence is a subjective quality difficult to measure reliably, unless special experimental procedures are employed [79, 86, 87, 94, 113, 114, 169, 50]. Experiments that make such measurements have often revealed systematic discrepancies between subjective confidence reports and experimentally measured accuracy [116, 74, 168, 90, 169, 114, 105]. These discrepancies have been occasionally interpreted as evidence for suboptimality of the decision-making process or for disparate processes for computing choice and confidence. Because our POMDP model enabled us to predict confidence from accuracy (see previous chapter), we explored if it could also explain five well-documented discrepancies between accuracy and confidence. Based on the model's success, we suggest that these discrepancies are neither anomalies of the decision-making process nor do they necessarily indicate a divergence of the neural mechanisms that compute choice and confidence. Rather, these phenomena are expected signatures of a decision-making process that infers the choice and its associated confidence in a unified framework.

Some discrepancies arise in an optimal decision-making process when the decision maker has incomplete knowledge about the environment and needs to resolve uncertainties about the reliability of observations. Others seem to exist from an experimenters' perspective because the exact information used by the subject is hidden to the experimenter. Our POMDP model, described

---

This chapter is mainly outsourced from [80].

in the previous chapter, explains commonly observed discrepancies between accuracy and confidence such as the "hard-easy" effect [74, 42, 133], higher confidence with increased variability of sensory observations despite reduction of accuracy [47, 169], different confidence ratings in simultaneous versus sequential reports of choice and confidence [87, 133, 154], discrepancy between sensitivities of accuracy and confidence (d' vs. meta-d') [92, 48], and the seemingly larger effect of choice-congruent observations on confidence reports [168, 114].

4.2 Hard-easy effect

The hard-easy effect, which has been documented extensively [74, 133], is the tendency to overestimate the likelihood of one's success for difficult decisions and underestimate it for easy decisions. In the face of uncertainty about the stimulus in a given trial, the model computes confidence across all possible stimuli (marginalization). However, when the experimenter measures accuracy for each stimulus strength, this marginalization does not occur as the experimenter knows the exact stimulus on each trial [40]. The model's uncertainty about the stimulus, therefore, causes overconfidence in difficult trials and underconfidence in easy ones.

As shown in Fig. 4.1a, the POMDP model predicts this hard-easy effect after marginalization over coherence. Since the model's Gaussian observation distribution closely approximates the true observation distribution (especially for the low coherence levels, (Fig 3.3a), it approximates well the confidence of the true generative model, as shown in figure 4.1b. However, the model does exhibit a small underconfidence bias since it considers the full range of continuous coherence levels. As expected, this bias is larger in the region where the coherence levels are further apart (and consequently the observations overlap less), which in our task are the easier trials (higher coherences; see monkey M1's plot), and for experiments with Monkey M2 where the task did not use the 1.6% coherence level. Overall, these results illustrate how task structure itself could create a bias in confidence for an optimal decision maker.

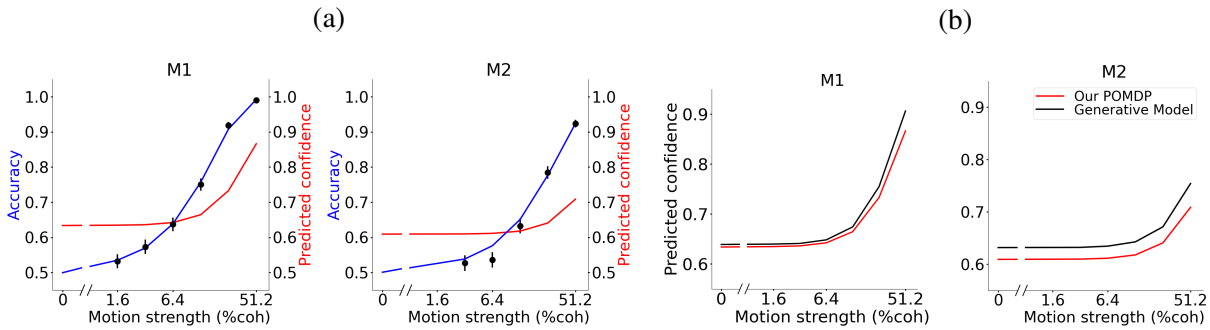


Figure 4.1: **POMDP model explains the hard-easy effect. a)** The hard-easy effect. Uncertainty about the strength of observed evidence makes the model more confident than warranted by accuracy on hard trials and less confident than warranted by accuracy on easy trials (compare blue and red curves). The psychometric and predicted confidence functions are adopted from Figures 3.5a and 3.5b. **b)** While the major reason for the hard-easy effect is marginalization over coherence by the decision maker, our POMDP model also predicts a small underconfidence bias (red curve) compared to using a generative model (black curve) that assumes the exact set of coherence levels in the experiment is given.

4.3 Opposing effects of the variability of observations on choice and confidence

A common observation in past studies has been that increasing the variability of the stimulus reduces subjects' accuracy but increases their confidence about their choices [119, 47, 169]. Our POMDP model shows that this seemingly paradoxical effect of stimulus variability arises naturally in an optimal inference framework when the subject does not have access to the true model of the environment (in this case, the true observation noise).

Stimulus variability effects have been explored in tasks where subjects were trained using a baseline (lower) stimulus variability, before being tested on higher variability. Further, trials with different levels of stimulus variability were randomly intermixed. Consequently, our model postulates that subjects continued to rely on the observation noise learned during initial training, and

used this model for choice and confidence in the high variability trials. Higher variability (larger w_z^2) decreases accuracy (Fig. 4.2a) by generating more overlapping observations for different motion directions.

Higher variability also generates extreme observations (far from the mean) more often, including ones in favour of the incorrect choice (e.g., negative coherence observations when the true coherence is positive). These extreme observations, although frequent in the high variability regime, are not expected based on the observation noise learned during training in a low variability regime. As a result, the POMDP model considers these extreme observations highly discriminative, resulting in a higher confidence with a concomitant decrease in the probability of choosing the sure-bet option when presented [169], especially in high and medium difficulty level trials (Figs. 4.2b and 4.2c).

To further understand this phenomenon, we adopted the intuitions and ideas suggested in previous work [119, 47, 169]. We explored the change in probability of rejecting the sure-bet option (indicating high confidence) in trials with a low coherence level c for a specific stimulus duration t and no observation cost. Suppose the true coherence is positive (the case where the coherence is negative is similar). The sum of observations comes from a Gaussian distribution with mean tc and variance tw_z^2. Choosing or rejecting the sure-bet option can be mapped to two thresholds on the sum of observations, one for each direction. This mapping depends on σ_0 and σ_z, and consequently w_z (indirectly).

Figure 4.2d shows the distribution of the sum of observations for two example stimuli with the same positive coherence level (+6.4%, green dotted line) and duration (250ms) but different variability, with low variability shown as the black Gaussian curve and high variability as the gray Gaussian curve. The plot also shows the sure-bet selection/rejection thresholds (black dotted lines) learned during training with the low variability curve for this example with +6.4% coherence. The low and high variability curves intersect each other at two points (red dotted lines). Note that the sure-bet selection/rejection thresholds (black dotted lines, fixed after training) are lesser than or greater than the intersection points (red dotted lines), implying that these learned thresholds are in the area where probability density for the higher variability stimulus (gray curve) is higher.

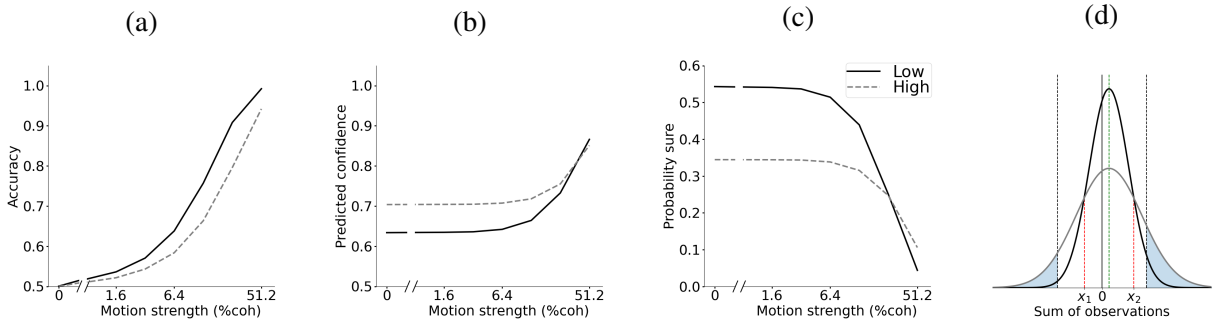


Figure 4.2: **Opposing effect stimulus variability on accuracy and confidence. a)** Sudden increase of stimulus variability following training with lower variability reduces accuracy **b)** This increase however boosts confidence **c)** Increase in variability also decreases probability of sure-bet rejection if presented. This dissociation occurs because the model relies on the observation noise learned during the lower-variability training to render choices on the higher variability trials. d) The increase in the probability of sure-bet rejection after a sudden increase in stimulus variability is illustrated here for two trials with the same coherence and duration but different stimulus variability. Distributions of the sum of observations for low (black Gaussian curve) and high variability (gray Gaussian curve) intersect at two points (red dotted lines). A high sure-bet rejection threshold on confidence (e.g., 85% in this example) learned during training with low variability stimuli maps to two thresholds (dotted black lines) on the sum of observations that fall outside of the intersection points. Given these fixed thresholds, the probability of sure-bet rejection is higher (larger blue filled areas under the curve) when stimulus variability is suddenly increased.

This means that the area under the curve beyond these thresholds (blue filled areas), equal to the probability of sure-bet rejection (indicating high confidence), is larger for the high variability stimulus than the low variability stimulus (narrower dark curve) used during training. These results illustrate how higher confidence can be generated when the stimulus becomes more noisy.

4.4 Discrepancy of sensitivity for accuracy and confidence

The POMDP model also explains experimentally observed differences between the sensitivity of accuracy and confidence to observations, commonly quantified with d' and meta-d', respectively [92]. d' and meta-d' are defined based on a signal detection theory (SDT) framework. d' quantifies the difference of sensory evidence distributions underlying the probability of correct and incorrect choices. However, meta-d' is related to the distribution of confidence ratings for those choices. For a binary confidence rating (low or high confidence, similar to rejecting or choosing the sure-bet option), meta-d' contrasts the probability of a high confidence rating for a correct response with that of an error. Some studies have reported that confidence ratings are not consistent with the sensitivity of the choice accuracy (d') [92, 49, 20, 107]. However, for an SDT ideal observer meta-d' and d' have to be similar in the absence of variability in the confidence rating threshold. Therefore, it has been suggested that the different meta-d' and d' in experimental data must be due to loss of information for confidence judgments or different neural mechanisms for confidence and choice [92, 48].

In the absence of an observation cost, where the POMDP model uses all available evidence, its d' and meta-d' match each other, similar to SDT. That would be true regardless of whether the decision maker does or does not have access to the exact model of the environment. However, if there is an early termination of information gathering, then meta-d' could diverge from d'. This discrepancy emerges in the model not because of distinct mechanisms for choice and confidence, but because early terminations of the decision-making process have quantitatively distinct effects on the choice accuracy and the likelihood of high confidence ratings for correct and incorrect choices. Because early terminations curtail the use of evidence, they reduce accuracy and, therefore, decrease d'. Further, in the face of uncertainty about the reliability of evidence, early terminations are associated with higher confidence (Figs. 3.4c and 4.3a). This combination means that for a wide range of model parameter values, the model makes more errors but it is also more confident about its choices compared to a model without an observation cost. Critically, the confidence is inflated more on error than correct trials (Fig. 4.3c, reducing meta-d'. This reduction could be sub-

stantially larger than the reduction of d'. Consequently, the model could generate meta-d' values smaller than its d', even though it computes the choice and confidence through the same optimal process.

Figure 4.3 illustrates these effects by simulating intermediate coherence (+12.8%) trials with 400 ms duration and subjecting the model choices and confidence to the d' and meta-d' calculations. Model parameters are inherited from Monkey M1 except for the addition of an observation cost (10^{-4}/observation). Early in the trial, observation noise can temporarily produce large positive or negative inferred μ_t, and thus high confidence (Fig. 4.3a, yellow lines illustrate $mean \pm 2 \times s.d.$ of the inferred μ_t). Such large μ_t are much less likely at later times because of the correction of excessive early confidence with additional observations. These later corrections, however, are prevented if the termination bounds (Fig. 4.3a, white lines) are reached earlier. Such occasional early terminations reduce the model accuracy by only 2% for this motion coherence (from 81% with no observation cost to 79%), but increase the overall probability of high confidence choices by 19% (from 65% to 84%) (Fig. 4.3b). The corrective effect of additional observations on confidence is more pronounced when the initial choice is incorrect as new observations are more likely to cancel the extreme noise that lead to early error choices. Consequently, early terminations increase the fraction of high confidence responses for incorrect choices by 39% (from 31% to 70%), whereas the increase for correct choices is 15% (from 72% to 87%) (Fig. 4.3c). This reduces the contrast of confidence for correct and error choices, resulting in a reduction of meta-d'. This reduction is larger than the very modest reduction of d', bringing the ratio of meta-d' to d' to 0.74, significantly below 1 (Fig. 4.3d).

The reduction of meta-d' could happen even when the overall confidence rating does not increase in the model, as meta-d' depends on the contrast of confidence for correct and error choices, which could be differentially affected by early terminations with or without an overall confidence increase. Generally, meta-d' to d' ratios below 1 are common for a wide range of POMDP model parameters matching a common result in past behavioral studies [48]. Further, the model predicts a mismatch between meta-d' and d' in reaction-time tasks, where the decision maker initiates a response as soon as reaching a decision. Overall, distinct d' and meta-d' values can arise in the

POMDP framework not because different information shapes choice and confidence, but rather when the decision-making process can stop due to a termination criterion without utilizing all the available information. This important alternative has been largely neglected in past explanations of mismatching meta-d' and d'.

4.5 Sensitivity of confidence measurements to simultaneous versus sequential reports of choice and confidence

The POMDP model can also be applied to reaction time tasks (besides fixed duration tasks), where subjects report their choice as soon as they are ready. In these tasks, the experimenter may ask for either a simultaneous or sequential report of choice and confidence [87, 133, 154]. Past experiments have shown that for simultaneous reports of choice and confidence, confidence for incorrect choices often increases with stimulus strength, compatible with the predictions of bounded accumulation models such as the drift diffusion model (DDM) [87, 154]. However, for sequential reports, confidence for incorrect choices decreases with stimulus strength, compatible with the predictions of signal detection theory [79, 133, 154].

POMDP predicts both patterns (Fig. 4.4a and 4.4b). In fact, being a normative model, POMDP goes beyond prediction and also explain why confidence increases with stimulus strength even in incorrect trials in the case of simultaneous reports of choice and confidence. Moreover, by focusing on the cause of trial termination in a reaction time task through utility maximization principle and comparing it with trial termination in a fixed-duration task, it also explains decrease of confidence with with stimulus strength in incorrect trials when the decision maker reports confidence after the choice.

The reason behind the increase of confidence with stimulus strength in simultaneous report of confidence and choice is quite similar to the mechanism explained in the previous section. Observation gathering terminations after a short period of time are associated with higher confidence (Fig. 3.4c). As pointed out in Fig.4.3a, this is especially important in incorrect trials with strong stimulus. When the stimulus strength is high, incorrect decisions after many observations are very unlikely. The reason behind an incorrect choice in an easy trial is early extreme observations which

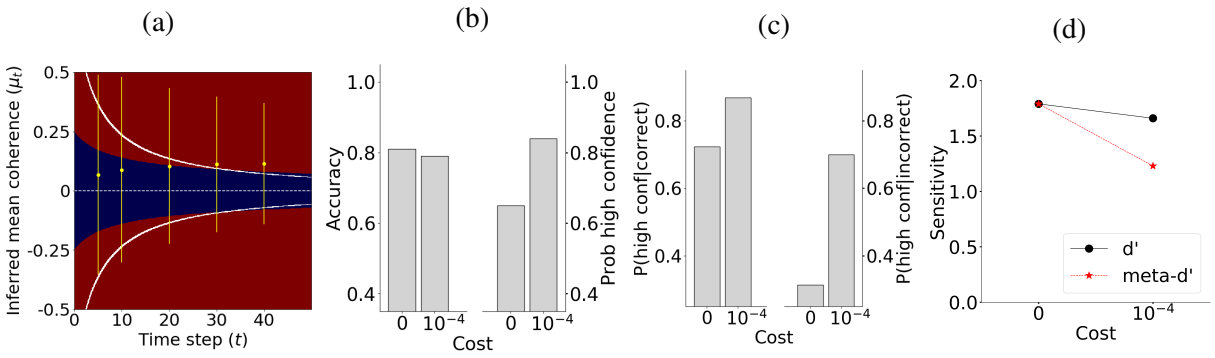


Figure 4.3: **POMDP explains different values for d' and meta-d'. a)** Observation noise could cause highly variable μ_t at the beginning of a trial, and thus temporarily produce excessive confidence. This excessive confidence may become permanent if the decision-making process is stopped by reaching the termination bounds. Solid white lines show the two decision termination bounds (observation cost, 10^{-4}). Thresholds for separating low and high confidence ratings are shown as boundaries between blue (low confidence) and red (high confidence) regions. The horizontal dashed line shows the boundary that separates right and left direction choices based on the sign of the inferred coherence. Yellow dots and lines show $mean \pm 2 \times s.d.$ (95% of the distribution mass) of the inferred coherence for a particular stimulus strength (c = +12.8%) at a few different time steps (10 ms per step). Temporary excessive confidence due to early termination is more prominent for the incorrect trials (negative μ_t in this simulation). **b)** Early termination can cause a modest reduction of accuracy and a marked increase of high-confidence ratings.**b)** Early termination can cause a larger increase in the probability of high confidence ratings for incorrect than correct choices. **c)** Changes of accuracy (c) and confidence ratings (d) can lead to a larger drop in meta-d' than d'. Model parameters are identical to those for monkey M1, except for the observation cost.

had not been canceled out by further observations due to hitting the termination bound. As a result, incorrect easier trials have much shorter duration and therefore higher confidence compared to hard

ones (Fig. 4.4a).

To explain the confidence pattern in sequential report of confidence and choice, we should look at the difference between fixed-duration and reaction time tasks. In a fixed duration task the subject commits to a choice early in the trial when the cost of gathering observation is higher than the increase in the expected utility. Importantly, the trial does not terminate with early commitment and the subject has to wait till the end of the trial to obtain the reward. In reaction time task, however, the decision maker controls the length of the trial which produces more incentive for commitment to a choice in addition to the cost of observation gathering. The decision maker gets the reward earlier if they select the correct choice earlier [65]. Moreover, the next trial and consequently a potential reward would become available sooner with earlier choice selection. For this reason, inter-trial interval and penalty time are accounted in modeling the cost function in decision making experiments [41].

Due to the bolder presence of extra factors described above in the cost function of reaction time task, even after selecting the choice, observations would be gathered if presented. In other words, observation cost is still worth the increase in the expected reward in the absence of other factors. The sensory and motor delays in the neural circuits underlying decisions usually amount to around 250 ms or more, lead to these extra observations. These observations do not contribute to the choice. They do not contribute to confidence report either if it is simultaneous with choice report. Sequential report of choice and confidence, however, opens up the possibility of revising confidence based on these last few sensory observations.

Confidence in incorrect trials is especially susceptible to such revisions. Importantly, in easy trials, these observations are very likely to be in favor of the real choice, and consequently opposite of early extreme observations that had led to an incorrect choice. As a result, they mostly decrease the confidence in incorrect trials (similar to no-cost case in previous section and Fig. 4.3a). In fact, since the easy incorrect trials are very short, the extra observations have a notable effect and might even lead to a change of mind about the true choice [125, 154], and consequently a confidence of less than .5 about the selected choice. These extra observations carry more information when the stimulus is stronger, and thus decrease the overall confidence about the incorrect choice more,

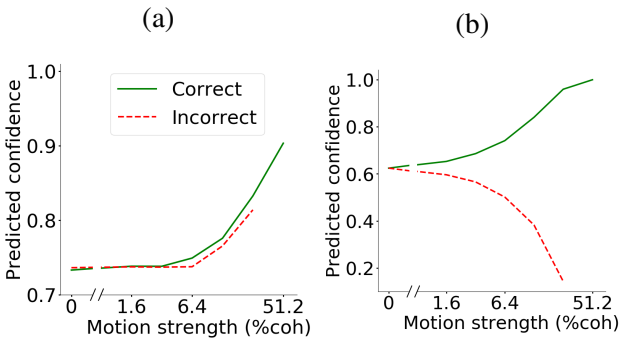


Figure 4.4: **POMDP explains different patterns of confidence report in RT task. a)** Simultaneous reports are commonly associated with higher confidence for erroneous choices on trials with stronger stimuli. This pattern is partly caused by lower decision times for stronger stimuli and the dependence of model confidence on elapsed time (Fig. 3.4c). **g)** In contrast, sequential reports of choice and confidence could be associated with reduced confidence for incorrect choices for stronger stimuli. This pattern is caused because sensory and motor delays render the last observations inconsequential for the choice but the model uses those observations to refine its confidence report following the choice.

resulting in decrease of overall confidence with stimulus strength in incorrect trials (Fig. 4.4b).

4.6 Effects of choice-congruent and choice-incongruent evidence

The last phenomenon we explore in this section is whether confidence reports are more strongly influenced by evidence congruent with the choice compared to incongruent evidence. Previous studies have reported that whereas choice is shaped by the balance of evidence for different options, confidence is more strongly shaped by choice-congruent evidence [168, 114]. These results have been interpreted as support for processes that compute confidence after the choice by readjusting the weight of evidence based on the choice (a form of confirmation bias). Our POMDP model demonstrates that this interpretation is not unique. Rather, existing experimental results could

be explained without assuming distinct choice and confidence processes, or choice-dependent re-weighting of evidence.

A key feature of analyzing data based on the POMDP framework is to distinguish the observations used by the subject and those analyzed by an experimenter who monitors the subject's behavior. Because the experimenter does not have access to the subject's observations as encoded in the nervous system, analysis of data has to rely on the expected distribution of evidence given stimulus properties (e.g., using filters on the stimulus [168]) or recordings from the brain (e.g., electrocorticography or ECoG [114]). Such estimated observations could markedly diverge from the actual observations used by the subject. A wide variety of mechanisms could underlie such a divergence, including decision bounds or other termination criteria unknown to the experimenter, sampling rates that mismatch the stimulus design, shifts in spatial or temporal attention during a trial, noise in the representation of sensory information by neural responses, or recording noise from the brain.

To clarify the significance of the divergence of observations used by the decision maker and those the experimenter uses to investigate behavior, consider the case where a decision maker uses only a proportion of the observations analyzed by the experimenter (n out of the total t samples, $n < t$). In this situation, the $t - n$ samples not used by the subject act as noise in the analyses. Classification of choice based on stimulus fluctuations reveals equal and opposing influence of stimuli supporting different alternatives as both used and unused observations come from the same distribution. However, conditional on the subject's choice, the proportion of choice-congruent observations is higher in the portion of the stimulus used by the subject, compared to the unused portion. This is simply because the sum of random variables drawn independently from the same distribution being positive is evidence in favor of each of these variables being positive. If we reorder the observations in a way that $z_1, \ldots, z_n$ become the ones used by the subject and $z_{n+1}, \ldots, z_t$ are the unused ones (only to simplify the equations), we have:

$$\sum_{j=1}^{n} z_j > 0 \rightarrow \forall\ 1 \leq i \leq n\ \&\ n+1 \leq l \leq t : P(z_i > 0) > P(z_l > 0) \tag{4.1}$$

And also:

$$\sum_{j=1}^{n} z_j > 0 \rightarrow \frac{1}{n}\mathbb{E}\left[\sum_{j=1}^{n} z_j\right] > \frac{1}{t-n}\mathbb{E}\left[\sum_{j=n+1}^{t} z_j\right] \tag{4.2}$$

The inequalities of Equations 4.1 and 4.2 have a profound side effect for quantification of the influence of individual observations on confidence. If we divide observations based on whether they support the choice, the ratio of total choice-congruent observations to total incongruent observations will be higher for the set of observations used by the decision maker (the n samples) than those used in the experimenter's analyses (all t samples). As a result, a classifier that uses all t observations to predict the decision maker's confidence has to give a larger weight to the choice-congruent observations to compensate for the dilution of congruent evidence caused by the unused stimulus samples.

To demonstrate this, we simulated a fixed-duration version of the random dots task with binary confidence ratings (low vs. high). For any stimulus strength and with $n < t$, a logistic classifier fit to the proportion of high confidence ratings by the POMDP model yielded larger weights for congruent than incongruent observations (Fig. 4.5a). In contrast, a logistic classifier fit to right and left choices based on stimulus fluctuations revealed equal and opposing weights for positive and negative samples as both used and unused observations come from the same distribution. As expected from Equation 4.2, the imbalance of the weights of the confidence was more pronounced for smaller n. To further demonstrate the inevitable imbalance of the weights, we compared the prediction accuracy of the confidence classifier with two alternative classifiers: one forced to have balanced weights for congruent and incongruent observations and a second classifier that had access only to the congruent evidence (Fig. 4.5b). Similar comparisons were used in past studies [114]. The confidence classifier with balanced weights had a lower prediction accuracy, especially for low n, where its accuracy was even lower than the classifier that totally ignored incongruent observations.

With similar reasoning, choice-congruent observations gain a higher weight in predicting confidence when the experimenter uses a subset of the observations used by the subject (Supplementary Fig. 1).

Although the example above focuses on a particular source of discrepancy between the observations used by the decision maker and experimenter (different number of used samples), the conclusion generalizes to other sources of discrepancy. Some of these sources such as neural noise are almost always present and quite difficult to correct the analyses for. Essentially, the observations analyzed by the experimenter are almost always noisy estimates of the observation used by the decision maker: $z_j^{experimenter} = z_j^{subject} + \zeta$, where ζ denotes noise with an often unknown magnitude. The neural noise causes the same dilution of choice congruent evidence explained in the example above. Consequently, the experimenter is bound to estimate a higher weight for congruent samples in the analyses even when $n = t$ and even though such weight imbalance may not exist for the decision maker. Large enough noise can even make a classifier that only uses choice-congruent observations better than a balanced classifier (Figs. 4.5c and 4.5d).

4.7 Discussion

Discrepancies between accuracy and confidence have been commonly considered as evidence for suboptimal decision-making or distinct processes that underlie choice and confidence. Our POMDP framework challenges these interpretations by showing that a normative Bayesian decision maker optimizing a reward function elicits the same discrepancies between confidence and accuracy as those identified in humans and experimental animals. We explored five common discrepancies in this chapter. Two of them arise from the decision maker's incomplete knowledge of the environment. The first one is the hard-easy effect, where decision makers are over-confident for difficult choice and under-confident for easy choices [42, 133]. This effect arises mainly from the decision maker's marginalization over the stimulus strength. Also, our model creates an extra small amount of underconfidence by estimating the true structure of the task by a continuous model. The second discrepancy is the opposing effects of stimulus variability on choice and confidence, where subjects become less accurate but paradoxically over-confident about more variable stimuli [169]. This effect arises from another form of incomplete knowledge about the environment: attribution of the observation noise learned in less variable environments to newly experienced, more variable conditions.

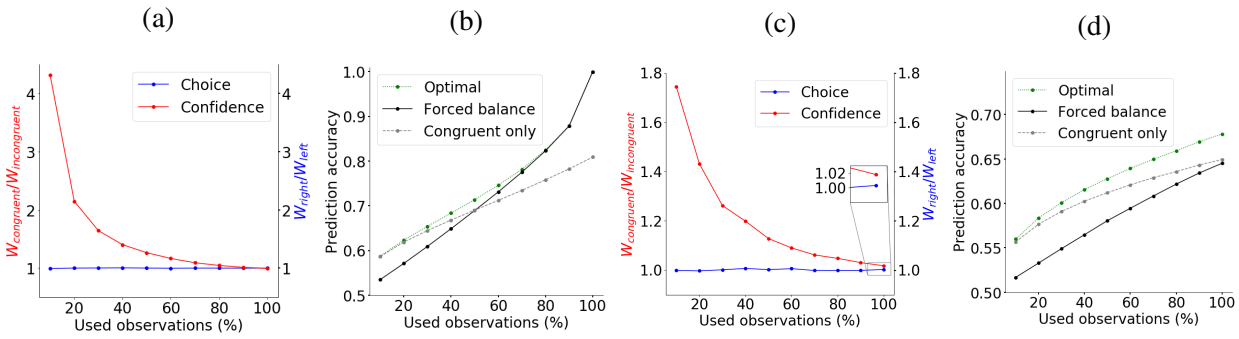


Figure 4.5: **POMDP explains seemingly higher influence of choice-congruent evidence on confidence. a)** We simulated a POMDP model that uses a fraction of observations available in a trial unbeknownst to the experimenter. The observations supporting opposing choices equally inform the model's behavior. However, an experimenter who uses a classifier to predict choices and confidence based on all observations in the trial finds an apparently larger influence of choice-congruent observations on confidence. **b)** Forcing the classifier to have balanced weights for all observations causes lower prediction accuracy of confidence ratings, especially when the proportion of used evidence is low.**c-d)** Same as a-b but observations accessible to the experimenter are noisy estimates of observations available to the decision maker. Such noise reduces the prediction accuracy of the experimenter's classifier, but more important, it also causes imbalanced weights in the optimal classifier (c) and lower performance of the balanced classifier (d). That is true even when both the decision maker and experimenter use all the available observations (inset in c). The noise in these simulations comes from a zero-mean Gaussian distribution with a variance of 25% larger than w_z^2.

The other three discrepancies between accuracy and confidence are explained by our model as arising from the experimenter's incomplete knowledge of the subjects' decision-making process. The third discrepancy is the inequality of d' and meta-d', which has attracted much attention lately as experimental support for distinct processes underlying choice and confidence [92]. We show that this difference could arise even when a unitary process shapes both choice and confidence, as

in our model. A cost-based termination criterion for the decision-making process could affect accuracy and confidence differently. Whereas the overall accuracy decreases due to early termination, confidence can increase especially for incorrect choices, causing unequal d' and d'. It is therefore impossible to uniquely interpret meta-d' in the absence of accurate knowledge about the form of the termination criterion. However, common task designs for measuring choice and confidence often preclude such knowledge.

It is also important to mention that in the presence of observation cost, meta-d' depends on the confidence rating threshold in the POMDP model. This sensitivity questions one of the key assumptions in the definition of meta-d' — independence of meta-d' from the confidence rating criterion — and cautions about interpretations of meta-d' results without knowing about the variability of confidence rating thresholds across subjects in an experiment.

The fourth discrepancy is based on the observation that confidence reports differ in experiments that interrogate confidence simultaneously with the choice [87] or after the choice [154]. This difference arises in our model because sequential reports of choice and confidence allow revising one based on information unused for the other. For example, when confidence reports follow the choice, sensory observations in the processing pipeline that were unavailable at the time of the choice could change confidence [154].

The fifth discrepancy that the model explains is the hypothesis that confidence is more strongly influenced by choice-congruent observations than choice-incongruent observations [168, 114]. Although these experimental results could indicate post-choice re-weighting of observations for calculation of confidence, they could also arise from the experimenter's incomplete or inaccurate knowledge of the exact observations used by the decision maker. Many factors could engender such inaccuracy, including neural noise, which is often inaccessible to the experimenter, device noise, which is difficult to eliminate for electrophysiological and imaging techniques, or termination criteria for the decision-making process, which the experimenter may be unaware of or unable to identify.

To clarify our conclusion, we do not imply that dual or hierarchical processes for choice and confidence could not exist. Nor do we exclude the possible existence of mechanisms that revise

confidence by post hoc choice-dependent re-weighting of the observations. Rather, we conclude that existing experimental results are insufficient to support such mechanisms as they are also compatible with simpler, more parsimonious mechanisms in which a unitary process underlies both choice and confidence. It is further illuminating that the unitary process explored in this chapter is based on POMDPs, a normative Bayesian framework based on expected reward maximization. In light of our POMDP model, existing experimental results should be carefully reconsidered and better experiments should be developed to test the necessity of more complex or disparate mechanisms for choice and confidence.

4.8 Methods

Simulations for increased stimulus variability

We used the POMDP model with parameters fit to Monkey M1's data. For the low variability regime, the standard deviation of observations was $w_z = .9$ and the learned standard deviation was $\sigma_z = 1.6$. For the high variability regime, the standard deviation of observations was $w_z = 1.5$ without changing σ_z or any other parameter in the POMDP model.

Exploring the effect of cost on sensitivity measurements and confidence report

We compared the POMDP model obtained from Monkey M1 with a model with similar parameters (w_z, σ_z, and σ_0) but with an observation cost of 10^{-4} added to establish decision termination bounds in trials with coherence of 12.8% and duration of 400ms. The confidence report was in the form of a binary rating (low or high) with a threshold of 0.63 applied to the belief about the choice. The sensitivity (d') and meta-d' were both 1.79 for zero observation cost. Increasing the cost to 10^{-4} decreased d' to 1.66 and meta-d' to 1.23. We used 1 million samples to ensure the results were robust. The code from [92] was used to calculate meta-d'. Slightly higher prior standard deviation ($\sigma_0 = .75$) was used for better visualization of the effect in Fig. 4.3a. Qualitatively similar results are obtained for other motion coherence levels and durations.

Simulations for simultaneous and sequential reports of choice and confidence

We used the following POMDP model parameters: $w_z = .4$, $\sigma_z = .75$, and $\sigma_0 = 5$ with the 7 discrete coherence levels used in our monkey experiment and a constant observation cost of 2×10^{-3} per 10 ms to simulate the reaction time task with 20,000 trials for each coherence. In the simultaneous report version of the model, both confidence and choice were calculated from the observations received prior to the model reaching its decision termination bounds. In the sequential report version, calculation of confidence continued to be influenced by observations during a 250 ms non-decision time after the choice.

Prediction power of choice-congruent and choice-incongruent observations

First, we simulated the random dots motion discrimination task with one coherence ($w_z = 1, c = 10.0\%$), one duration (800 ms) and a binary confidence rating of low or high with a POMDP model that had an exact model of the world (i.e., with the true w_z and c) but used the first n observations out of $t = 80$ (step size = 10 ms) observations. For each n, the confidence threshold was set to a value that made the probability of high confidence around 0.5.

To generate data points for Figures 4.5a and 4.5b, we trained logistic regression classifiers to predict the simulated choices and confidence ratings. 10 million trials were simulated for these analyses to ensure robust and accurate results. Our classifiers were implemented using the `scikit-learn` Python library [112]. For choice, the features of our classifier were the sum of positive observations and the sum of negative observations throughout each trial, including those beyond the first n samples used for simulating choice and confidence. For confidence, the features were the sum of choice-congruent observations and the sum of choice-incongruent observations throughout each trial. For the balanced classifier, to ensure balance of weights, we used a classifier with a feature consisting of the sum of all observations signed according to the choice (positive for choice-congruent and negative for choice-incongruent) as one feature.

Chapter 5

BELIEF-BASED GROUP DECISION MAKING

5.1 Introduction

The importance of social decision making in human behavior has spawned a large body of research in social neuroscience and decision making [134, 73]. Human behavior relies heavily on predicting future states of the environment under uncertainty and choosing appropriate actions to achieve a goal. In a social context, the degree of uncertainty about the possible outcomes increases drastically as the behavior of others is much less predictable than the physics of the environment.

One approach to handling uncertainty in social settings is to act based on a belief about others. This approach includes inferring the consequences of one's own behavior under uncertainty as opposed to "belief-free" models [97] that simply select the action that has been rewarding in the past, given current observations [18, 51]. The difference between "belief-based" and "belief-free" models in social decision making is closely related to "model-based" and "model-free" approaches [34, 32] in non-social decision making, but with a greater emphasis on uncertainty due to the greater unpredictability of human behavior in social tasks.

In belief-based decision making, the subject learns a model of the environment, updates the model based on observations and rewards, and chooses actions based on a probabilistic "belief" about the current state of the world [33, 29, 51]. As a result, the relationship of the current action with rewards received and current observations is indirect. Besides the history of rewards received and the current observation, the learned model can also include other factors such as potential future rewards and more general rules about the environment. Therefore, the belief-based (model-based) approach is more flexible than belief-free (model-free) decision making [39, 38]. However, belief-based decision making requires more cognitive resources, for example, for simulation of future

---

This chapter is mainly outsourced from [83].

events. Thus, there is an inherent trade-off between the two types of approaches, and determining which approach humans adopt for different situations is an important open area of research [21].

Several studies have presented evidence in favor of the belief-based approach by quantifying the similarity between probabilistic model-based methods and human behavior when the subject interacts with or reasons about another human [166, 163, 99, 100, 69, 51, 5]. Compared to reasoning about a single person, decision making in a group with a large number of members can get complicated. On the one hand, having more group members disproportionately increases the cognitive cost of tracking minds compared to the cost of tracking the reward history of each action given the current observations. On the other hand, consistent with the importance that human society places on group decisions, a belief-based approach might be the optimal strategy.

How does one extend a belief-based approach for reasoning about a single person to the case of decision making within a large group? Group decision making becomes even more challenging when the actions of others in the group are anonymous (e.g., voting as part of a jury) [144, 110]. In such situations, reasoning about the state of mind of individual group members is not possible but the dynamics of group decisions do depend on each individual's actions.

To investigate these complexities that arise in group decision making, we focused on the Volunteer's Dilemma task, wherein a few individuals endure some costs to benefit the whole group [37]. Examples of the task include guarding duty, blood donation, and stepping forward to stop an act of violence in a public place [31]. In order to mimic the Volunteer's Dilemma in a laboratory setting, we use the thresholded binary version of a multiround Public Goods Game (PGG) where the actions of each individual are hidden from others [37, 2].

Using an optimal Bayesian framework based on Partially Observable Markov Decision Processes (POMDPs) [75], we propose that in group decision making, humans simulate the "mind of the group" by modeling an *average group member's mind* when making their current choices. Our model incorporates prior knowledge, current observations, and a simulation of the future based on the current actions for modeling human decisions within a group. We compared our model to a model-free reinforcement learning approach based on the reward history of each action as well as a previous descriptive method for fitting human behavior in the Public Goods Game. Our model

predicts human behavior significantly better than the model-free reinforcement learning and descriptive approaches. Furthermore, by leveraging the interpretable nature of our model, we are able to show a potential underlying computational mechanism for the group decision making process.

5.2 Human behavior in a binary public goods game

The participants were 29 adults (mean age 22.97 years old $\pm$ 0.37, 14 women). We analyzed the behavioral data of 12 Public Goods Games (PGGs) in which participants played 15 rounds of the game within the same group of N players ($N = 5$).

At the beginning of each round, 1 monetary unit (MU) was endowed (E) to each player. In each round, a player could choose between two options: *contribute* or *free-ride*. Contribution had a cost of C = 1 MU, implying that the player could choose between keeping their initial endowment or giving it up. In contrast to the classical PGG where the group reward is a linear function of total contributions [45], in our PGG, public goods were produced as a group reward (G = 2 MU to each player) if and only if at least k players each contributed 1 MU. k was set to 2 or 4 randomly for each session and conveyed to group members before the start of the session. The resultant amount after one round is therefore E - C + G = 2 MU for the contributor and E + G = 3 MU for the free-rider when public goods were produced (the round was a SUCCESS). On the other hand, the contributor has E - C = 0 MU and the free-rider has E = 1 MU when no public goods were produced (the round was a FAILURE).

Figure 5.1 depicts one round of the PGG task. After the subject acts, the total number of contributions, free rides, and the overall outcome of the round is revealed (success or failure in securing the 2 MU group reward) but each individual player's actions remained unknown. Additionally, as shown in the figure, the value of k for the current session was always presented on the screen to insure that the subjects had it in mind when making decisions. Although subjects were told that they were playing with other humans, in reality, they were playing with a computer that generated the actions of all the other $N-1=4$ players using an algorithm based on human data (see Methods). In each session, the subject played with a different group of players.

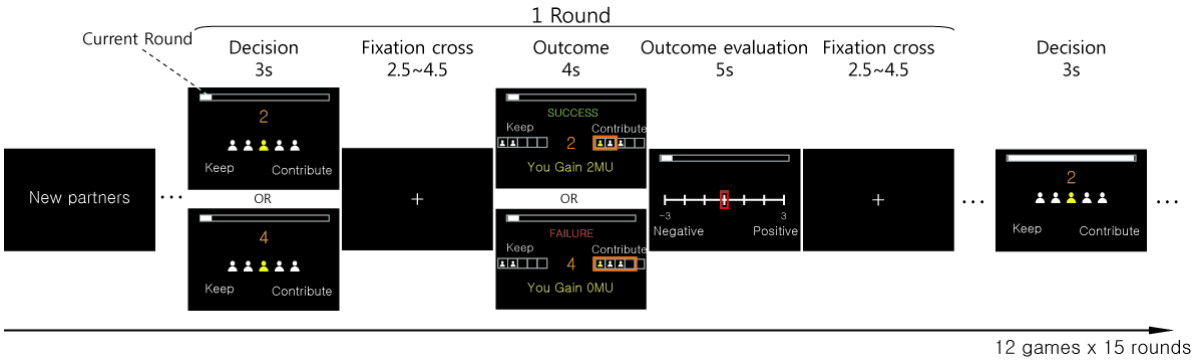


Figure 5.1: **Multi-round public goods game.** The figure depicts the sequence of computer screens a subject sees in one round of the PGG. The subject is assigned 4 other players as partners and each round requires the subject to make a decision: Keep 1 monetary unit (i.e., free-ride) or contribute. The subject knows whether the threshold to generate public goods (reward of 2 MU for each player) is 2 or 4 contributions (from the 5 players). After the subject acts, the total number of contributions and overall outcome of the round (success or failure) are revealed.

As shown in Figure 5.2a, subjects contributed significantly more when the number of required volunteers was higher with an average contribution rate of 55% ($SD = .31$) for $k = 4$ in comparison to 33% ($SD = .18$) for $k = 2$ (two-tailed paired sample t-test, $t(28) = 3.94$, $p = 5.0 \times 10^{-4}$, 95% CI difference = $[0.11, 0.33]$). In addition, Figure 5.2b shows that the probability of generating public good was significantly higher when $k = 2$ with a success rate of 87% ($SD = 0.09$) compared to 36% ($SD = .29$) when $k = 4$ (two-tailed paired sample t-test, $t(28) = 10.08$, $p = 8.0 \times 10^{-11}$, 95% CI difference =$[0.40, 0.60]$) (Figure 5.2b). All but 6 of the subjects contributed more when $k = 4$ (Figure 5.2c). Of these 6 players, 5 chose to free-ride more than 95% of the time. Also, success rate was higher when $k = 2$ for all players (Figure 5.2d).

The contribution rate of the subjects dropped during the course of the trial on average, especially for $k = 4$, but remained above zero. Figure 5.2e shows the average contribution rate across all subjects as a function of round number (1 to 15). We also compared the average contribution for the first five rounds with that for the last five rounds. For $k = 4$, the average contribution

probability across all subjects for the first five rounds was .6 ($SD = .20$) and significantly higher than that for the last five rounds (average across subjects = .49, $SD = .19$) (two-tailed paired sample t-test, $t(28) = 3.65$, $p = 0.001$, 95% CI difference =$[0.05, 0.17]$). For $k = 2$, the difference between the first five rounds (average = .53 , $SD = .32$) and the last five rounds (average = .50 , $SD = .33$) was insignificant (two-tailed paired sample t-test, $t(28) = 1.51$, $p = .14$, 95% CI difference =$[-0.01, 0.06]$).

The average contribution probability did not change significantly as subjects played more games (Figure 5.2f). In each condition, most of the players played at least 5 games (27 players for $k = 2$ and 26 for $k = 4$). For $k = 2$, in their first game, the average contribution rate of players was .37 ($SD = .25$) while in their fifth game, it was .30 ($SD = .24$) (two-tailed paired sample t-test, $t(26) = 1.34$, $p = 0.19$, 95% CI difference =$[-0.03, 0.17]$). When $k = 4$, the average contribution rate was .57 ($SD = .30$) in the first game and .61 ($SD = .35$) in the fifth game (two-tailed paired sample t-test, $t(25) = -0.69$, $p = 0.50$, 95% CI difference =$[-0.16, 0.08]$).

5.3 Probabilistic model of theory of mind for the group in the public goods game

Consider one round of the PGG task. A player can be expected to choose an action (*contribute* or *free ride*) based on the number of contributions they anticipate the others to make in that round. Since the actions of individual players remain unknown through the game, the only observable parameter is the total number of contributions. One can therefore model this situation using a single random variable θ, denoting the average probability of contribution by any group member. With this definition, the total number of contributions by all the other members of the group can be expressed as a binomial distribution. Specifically, if θ is the probability of contribution by each group member, the probability of observing m contributions from the $N-1$ others in a group of N people is:

$$P(m|\theta) = \binom{N-1}{m} \theta^m (1-\theta)^{N-1-m}. \tag{5.1}$$

Using this probability, a player can calculate the expected number of contributions from the others, compare it with k and decide whether to contribute or free-ride accordingly. For example,

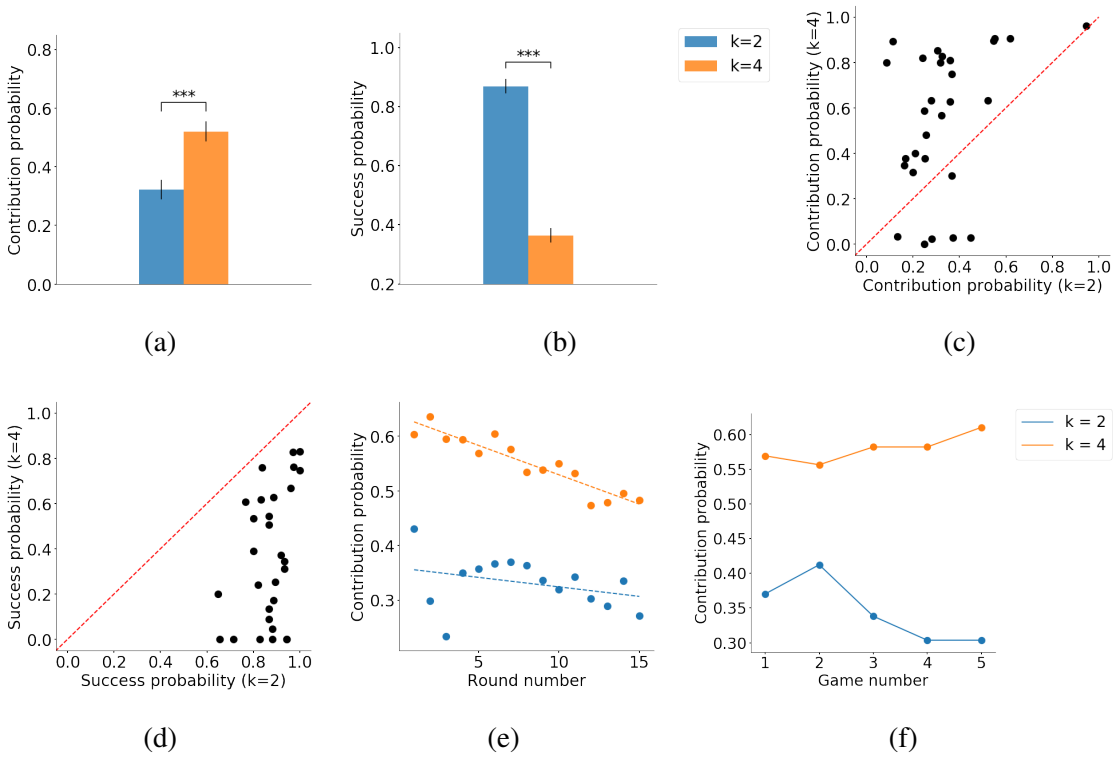


Figure 5.2: **Human behavior in the PGG task** (a) The average contribution probability across subjects is significantly higher when the task requires more volunteers (k) to generate the group reward. (b) Average probability of success across all subjects in generating the group reward is significantly higher when k is lower. Error bars indicate within-subject standard error [28]. (c) Average probability of contribution for each subject for $k = 2$ versus $k = 4$. Each point represents a subject. Subjects tend to contribute more often when the task requires more volunteers. (d) Average success rate for each subject was higher for $k = 2$ versus $k = 4$. (e) The average probability of contribution across subjects decreases throughout a game, especially for $k = 4$. Dotted lines are linear functions showing this trend for each k. (f) Average contribution probability across subjects as a function of number of games played. The contribution probability does not change significantly as subjects play more games.

if θ is very low, there is not a high probability of observing $k-1$ contributions by the others, implying free riding is the best strategy.

There are two important facts that make this decision making more complex. First, the player does not know θ. θ must be estimated from the behavior of the group members. Second, other group members have a theory of mind as well. Therefore, they can be expected to change their strategy based on the actions of others. In fact, due to this ability in other group members, each player needs to simulate the effect of their action on the group's behavior in the future.

To model the uncertainty in θ, we assume that a probability distribution over θ is maintained in the player's mind, representing their belief about the cooperativeness of the group. Each player starts with an initial probability distribution, called the prior belief about θ and updates this belief over successive rounds based on the actions of the others. The prior belief may be based on the prior life experience of the player, or what they believe others would do through fictitious play [16]. For example, the player may start with a prior belief that the group will be a cooperative one but change this belief after observing low numbers of contributions by the others. Such belief updates can be performed using Bayes' rule to invert the probabilistic relationship between θ and the number of contributions given by Equation 5.1.

A suitable prior probability distribution for estimating the parameter θ of a Binomial distribution is the Beta distribution, which is itself determined by two (hyper) parameters α and β:

$$\begin{gathered} \theta \sim Beta(\alpha,\beta). \\ Beta(\alpha,\beta) : P(x|\alpha,\beta) \propto x^{\alpha-1}(1-x)^{\beta-1}. \end{gathered} \tag{5.2}$$

Starting with a prior probability $Beta(\alpha_1,\beta_1)$ for θ, the player updates their belief about θ after observing the number of contributions from the others in each round through Bayes' rule. This updated belief is called the posterior probability of θ. The posterior probability of θ in each round serves as the prior for the next round.

In economics, the ability to infer the belief of others is sometimes called sophistication [27, 35]. Here we consider a simple form of sophistication: we assume that each player thinks other group members have the same model as themselves (α and β). This is justifiable due to computational

efficiency and more importantly anonymity of players. As a result, with a prior of $Beta(\alpha_t, \beta_t)$ after observing c contributions (including one's own when made) in round t, the posterior probability of θ for the subject becomes $Beta(\alpha_{t+1}, \beta_{t+1})$ where $\alpha_{t+1} = \alpha_t + c$ and $\beta_{t+1} = \beta_t + N - c$.[1] Note that we include one's own action in the update of the belief because one's own action can change the future contribution level of the others.

Intuitively, α represents the number of contributions made thus far and β the number of free rides. α_1 and β_1 (that define prior belief) represent the player's a priori expectation about the relative number of contributions versus free-rides respectively before the session begins. For example, when α_1 is larger than β_1, the player starts the task with the belief that people will contribute more than free ride. Large values of α_1 and β_1 imply that the subject thinks the average contribution probability will not change significantly after one round of the game when updated with the relatively small number c as above.

Decision making in the PGG task is also made complex by the fact that the actual cooperativeness of the group itself (not just the player's belief about it) may change from one round to the next: players observe the contributions of the others and may change their own strategy for the next round. For example, players may start the game making contributions but change their strategy to free riding if they observe a large number of contributions by the others. We model this phenomenon using a parameter $0 \leq \gamma \leq 1$, which serves as a decay rate: the prior probability for round t is modeled as $Beta(\gamma\alpha_t, \gamma\beta_t)$, which allows recent observations about the contributions of other players to be given more importance than observations from the more distant past. Thus, in a round with c total contributions (including the subject's own contribution when made), the subject's belief about the cooperativeness of the group as a whole changes from $Beta(\alpha_t, \beta_t)$ to $Beta(\alpha_{t+1}, \beta_{t+1})$ where $\alpha_{t+1} = \gamma\alpha_t + c$ and $\beta_{t+1} = \gamma\beta_t + N - c$.

[1] Technically, this follows because the beta distribution is conjugate to the binomial distribution [101].

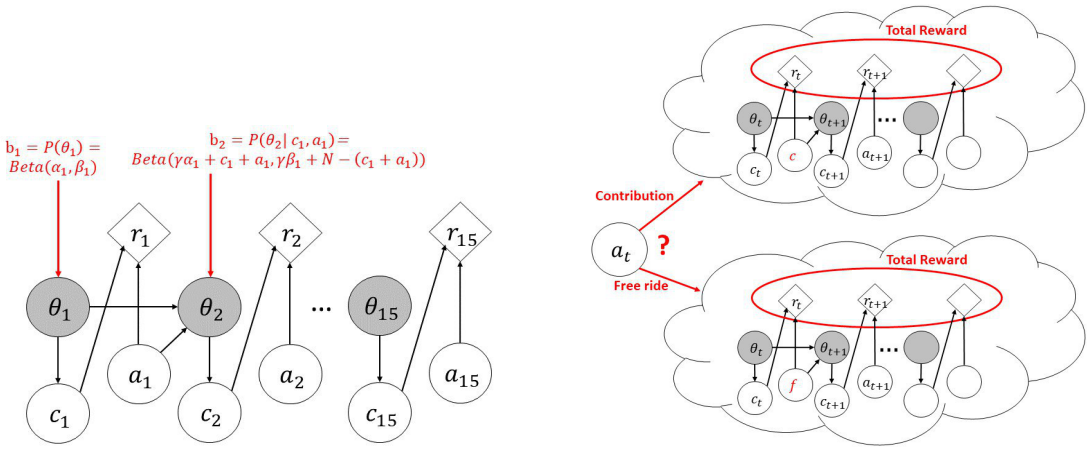
5.3.1 Action selection

How should a player decide whether to contribute or free ride in each round? One possible strategy is to maximize the reward for the current round by calculating the expected number of contributions by the others based on the current belief. Using Equation 5.1 and the prior probability distribution over θ, the probability of seeing m contributions by the others when the belief about the cooperativeness of the group is $Beta(\alpha,\beta)$ is given by:

$$\begin{aligned} P(m|\alpha,\beta) &= \int_0^1 P(m|\theta)P(\theta|\alpha,\beta)d\theta \propto \int_0^1 \binom{N-1}{m}\theta^m(1-\theta)^{N-1-m}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\ &\propto \binom{N-1}{m}\int_0^1 \theta^{\alpha+m-1}(1-\theta)^{\beta+N-m-2}d\theta \end{aligned} \tag{5.3}$$

One can calculate the expected reward for the *contribute* versus *free-ride* actions in the current round based on the above equation. Maximizing this reward however is not the best strategy. As alluded to earlier, the actions of each player can change the behavior of other group members in future rounds. Specifically, our model assumes that its own contribution in the current round increases the average contribution rate of the group in the future rounds. Equation 5.5 in Methods shows the exact assumptions of our model (with updates of $\alpha_{t+1} = \gamma\alpha_t + c$ and $\beta_{t+1} = \gamma\beta_t + N - c$ for its belief) about the dynamics of the actual (hidden) state of the environment. The optimal strategy therefore is to calculate the cooperativeness of the group through to the end of the session and consider the reward over all future rounds in the session before selecting the current action. Thus, an optimal agent would contribute for two reasons. First, contributing could enable the group to reach at least k volunteers in the current round. Second and more importantly, contributing encourages other members to contribute in future rounds. Specifically, a contribution by the subject increases the average contribution rate for the next round by increasing α in the next round (See the transition function in Methods).

Long-term reward maximization (as discussed above) based on probabilistic inference of hidden state in an environment (here, θ, the probability of contribution of group members) can be modeled using the framework of Partially Observable Markov Decision Processes (POMDPs) [75]. Figure 5.3a shows a schematic of the PGG experiment modeled using a POMDP and figure 5.3b



(a) (b)

Figure 5.3: **POMDP model of the multi-round public goods game.** (a) Model: The subject does not know the average probability of contribution of the group. The POMDP model assumes the subject maintains a probability distribution (”belief”, denoted by b_t) about the group’s average probability of contribution (denoted by θ_t) and updates this belief after observing the outcome c_t (contribution by others) in each round. (b) Action Selection: The POMDP model chooses an action (a_t) that maximizes the expected total reward ($\sum r_i$) across all rounds based on the current belief and the consequence of the action (contribute ”c” or free-ride ”f”) on group behavior in future rounds.

illustrates the mechanism of action selection in this model.

As an example of the POMDP model’s ability to select actions for the PGG task, Figures 5.4a and 5.4b show the best actions for a given round (here, round 9) as prescribed by the POMDP model for $k = 2$ and $k = 4$ respectively (the number of minimum volunteers needed). The best actions are shown as a function of different belief states the subject may have, expressed in terms of the different values possible for belief parameters α_t and β_t. This mapping from belief to actions is called a *policy*.

Our simulations using the POMDP model showed that considering a much longer horizon (e.g,

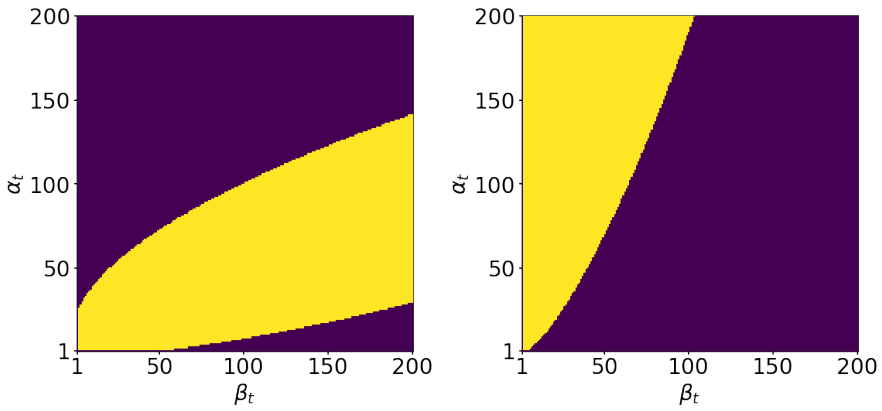


Figure 5.4: **Optimal actions prescribed by the POMDP policy as a function of belief state.** Plot (a) shows the policy for $k = 2$ and plot (b) for $k = 4$. The purple regions represent those belief states (defined by α_t and β_t) for which free-riding is the optimal action; the yellow regions represent belief states for which the optimal action is contributing. These plots confirm that the optimal policy depends highly on k, the number of required volunteers. For the two plots, the decay rate was 1 and t was 9.

50 rounds) instead of just 15 rounds gave a better fit to the subjects' behavior, suggesting that human subjects may be inclined to employ long horizons for group decision making tasks (see Discussion). Such a long horizon for determining the optimal policy makes the model similar to an infinite horizon POMDP model [149]. As a result, the optimal policy for all rounds in our model is very similar to the policy for round 9 shown in Figures 5.4a and 5.4b.

In summary, the POMDP model performs two computations simultaneously. The first computation is probabilistic estimation of the (hidden) average contribution rate through belief updates. The average contribution rate changes during the course of the game as players interact with each other. The second computation involves selecting actions to influence this average contribution rate and to maximize total expected reward. This is the action selection component, which is performed by backward reasoning from the last round.

5.4 POMDP model predicts human behavior in volunteer's dilemma task

The POMDP model has three parameters, α_1, β_1, and γ which determine the subject's actions and belief in each round. We fit these parameters to the subject's actions by minimizing the error, i.e. the difference between the POMDP model's predicted action and the subject's action in each round. The average percentage error across all rounds is then the percentage of rounds that the model incorrectly predicts (*contribute* instead of *free-ride* or vice versa). We defined accuracy as the percentage of the rounds that the model predicts correctly.

We also calculated the leave-one-out cross validated (LOOCV) accuracy of our fits [101] where each "left out" data point is one whole game and the parameters were fit to the other 11 games of the subject. It is important to note that our LOOCV accuracy is a *prediction* of the subject's behaviour in a game without any parameter tuning based on this game. Also, while different rounds of each game are highly correlated, the games of each subject are independent from each other (given the parameters of that subject) as the other group members change in each game.

We found that the POMDP model had an average fitting accuracy across subjects of 84% ($SD = 0.06$) while the average LOOCV accuracy was 77% ($SD = 0.08$). Figure 5.5a compares the average fitting and LOOCV accuracies of the POMDP model with two other models. The first is a "model-free" reinforcement learning model known as Q-learning: actions are chosen based on their rewards in previous rounds [152] with the utility of group reward, initial values, and learning rate as free parameters (5 parameters per subject – see Methods).

The average fitting accuracy of the Q-learning model was 79% ($SD = .07$) which is significantly worse than the POMDP model's fitting accuracy given above (two-tailed paired t-test, $t(28) = -6.75$, $p = 2.52 \times 10^{-7}$, 95% CI difference = $[-0.06, -0.03]$). Also, the average LOOCV accuracy of the POMDP model was significantly higher than the average LOOCV accuracy of Q-learning, which was 73% ($SD = .09$) (two-tailed paired t-test, $t(28) = 2.20$, $p = 0.037$, 95% CI difference =$[0.004, 0.08]$).

We additionally tested a previously explored descriptive model in PGG literature known as the linear two-factor model [162], which predicts the current action of each player based on the

player's own action and contributions by the others in the previous round (this model has three free parameters per subject – see Methods). The average fitting accuracy of the two-factor model was 78% ($SD = 0.09$) which is significantly lower than the POMDP model's fitting accuracy (two-tailed paired t-test, $t(28) = -4.86$, $p = 4.1 \times 10^{-5}$, .95% CI difference = $[-0.08, -0.03]$. Moreover, the LOOCV accuracy of the two-factor model was 47% ($SD = 20$), significantly lower than the POMDP model (two-tailed paired t-test, $t(28) = -7.61$, $p = 2.7 \times 10^{-8}$, 95% CI difference = $[-0.38, -0.22]$). The main reason for this result, especially the lower LOOCV accuracy, is that group success also depends on the required number of volunteers (k). This value is automatically incorporated in the POMDP's calculation of expected reward. Also, reinforcement learning works directly with rewards and therefore does not need explicit knowledge of k (however, a separate parameter for each k is needed in the initial value function for Q-learning – see Methods). Given that the number of free parameters for the descriptive and model-free approaches is greater than or equal to the number of free parameters in the POMDP model, the higher accuracy of POMDP is notable in terms of model comparison.

We tested the POMDP model's predictions of contribution probability for each subject for the two k values with experimental data (same data as in Figure 5.2c, see Methods). As shown in figures 5.5b and 5.5c, the POMDP model's predictions match the pattern of distribution of actual data from the experiments.

The POMDP model, when fit to a subject's actions, can also explain other events during the PGG task in contrast to the other models described above. For example, based on equation 5.3 and the action chosen by the POMDP model, one can predict the subject's belief about the probability of success in the current round. This prediction cannot be directly validated but it can be compared to actual success. If we consider actual success as the ground truth, the average accuracy of the POMDP model's prediction of success probability across subjects was 71% ($SD = .07$). Moreover, the predictions matched the pattern of success rate data from the experiment (Figures 5.5e and 5.5d). The other models presented above are not capable of making such a prediction.

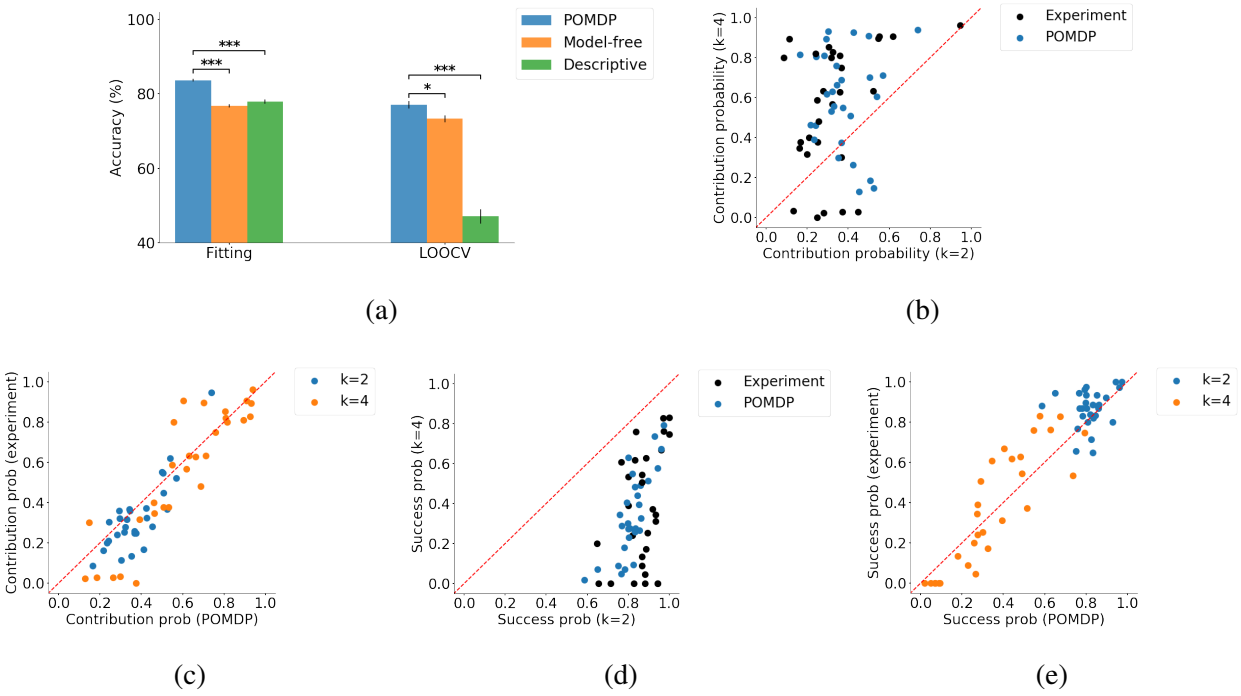


Figure 5.5: **POMDP model's performance and predictions about volunteer's dilemma task** (a) Average and LOOCV accuracy across all models. The POMDP model has significantly higher accuracy compared to the other models ($p < .05$ one star, $p < .01$ two stars, and $p < .001$ three stars). Error bars indicate within-subject standard error [28]. (b) POMDP model's prediction of a subject's probability of contribution compared to experimental data for the two k values (black circles, same data as in figure 5.2c). (c) Same data as (b) but POMDP model's prediction and the experimental data are shown for each k separately (blue for $k = 2$ and orange for $k = 4$) (d) POMDP model's prediction (blue circles) of a subject's belief about group success in each round (on average) compared to actual data (black circles, same data as in figure 5.2d). (e) Same data as (d) but POMDP model's prediction and actual data are shown for each k separately (blue for $k = 2$ and orange for $k = 4$).

5.5 Distribution of POMDP parameters

We can gain insights into the subject's behavior by interpreting the parameters of our POMDP model in the context of the task. As alluded to above, the prior parameters α_1 and β_1 represent the subject's prior expectations of contributions and free-rides respectively. Therefore, the ratio α_1/β_1 characterizes the subject's expectation of contributions by group members while the average of these parameters, $(\alpha_1+\beta_1)/2$, indicates the weight the subject gives to prior experience with similar groups before the start of the game. The decay rate γ determines the weight given to past observations compared to new ones: the smaller the decay rate, the more weight the subject gives to new observations.

We examined the distribution of these parameter values for our subjects after fitting the POMDP model to their behavior (Figures 5.6a and 5.6b). The ratio α_1/β_1 was in the reasonable range of .5 to 2 for almost all subjects (Figure 5.6c; in our algorithm the ratio can be as high as 200 or as low as $1/200$ – see Methods). The value of $(\alpha_1+\beta_1)/2$ across subjects was mostly between 40 (analogous to $40/(N/2)=16$ rounds of the game) to 120 (Figure 5.6d), suggesting that prior belief about groups did have a significant role in players' strategy, but it was not the only factor since observations over multiple rounds can still alter this initial belief. To confirm the effect of actions during the game, we performed a comparison with a POMDP model that does not update α and β over time and only uses its prior. The accuracy of this modified POMDP model was 66% ($SD=.17$), significantly lower than our original model (two-tailed paired t-test, $t(28)=-5.47$, $p=7.64\times10^{-6}$, 95% CI difference =$[-0.23,-0.11]$).

We also calculated the expected value of contribution by the others in the first round, which is between 0 and $N-1=4$ based on the values of α_1 and β_1 for the subjects. For almost all subjects, this expected value was between 2 and 3 (Figure 5.6e).

In addition, we calculated each subject's prior belief about group success (probability of success in the first round) based on α_1, β_1, and the subject's POMDP policy in the first round. As group success depends on the required number of volunteers (k), probability of success is different for $k=2$ and $k=4$ even with the same α_1 and β_1. Figures 5.6f and 5.6g show the distribution of

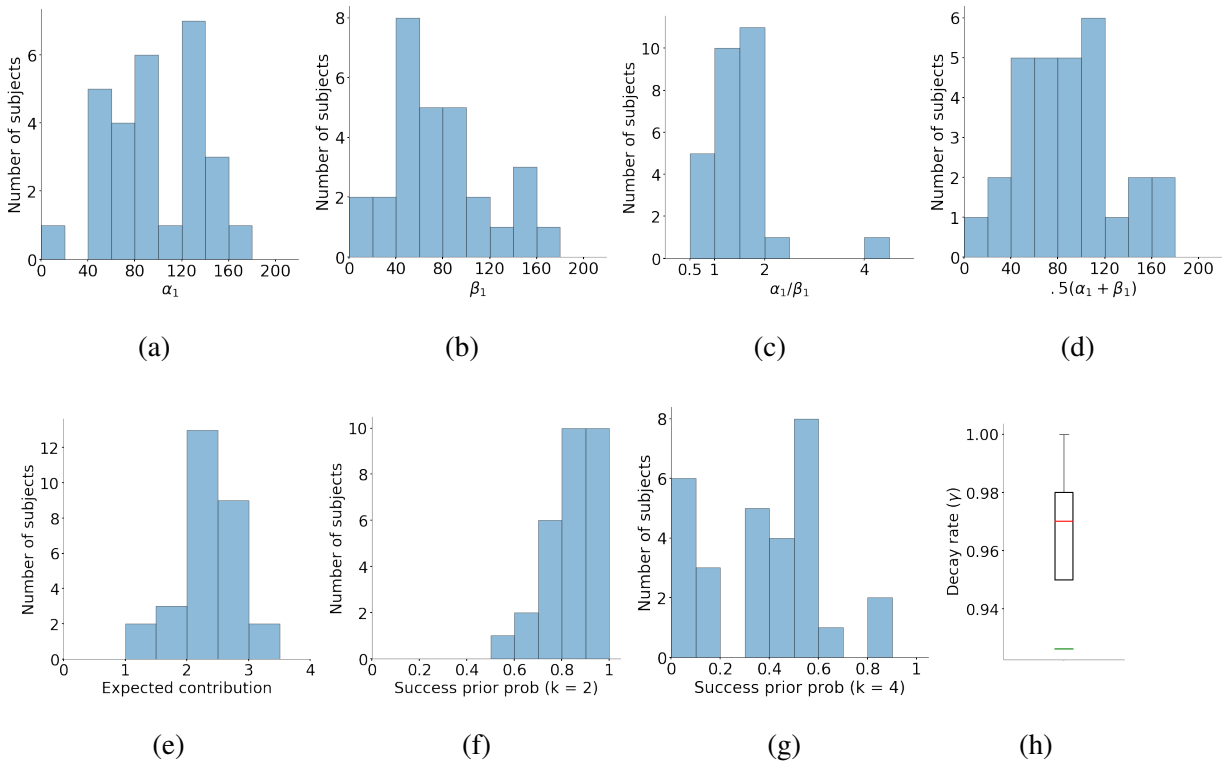


Figure 5.6: **Distribution of POMDP parameters across subjects.** (a) Histogram of α_1 across all subjects. (b) Histogram of β_1 across all subjects. (c) Histogram of the ratio α_1/β_1 shows a value between .5 and 2 for almost all subjects. (d) Histogram of $(\alpha_1+\beta_1)/2$. For the majority of subjects, this value is between 40 to 120. (e) Histogram of prior belief $Beta(\alpha_1,\beta_1)$ translated into expected contribution by the others in the first round. Note that the values, after fitting to the subjects' behavior, are mostly between 2 and 3. (f) When $k=2$, all subjects expected a high probability of group success in the first round (before making any observations about the group). (g) When $k=4$, almost all subjects assigned a chance of less than 60% to group success in the first round. (h) The box plot of decay rate γ across subjects shows that this value is almost always above .95. The median is .97 (orange line) and the mean is .93 (green line).

this prior probability of success across all subjects for $k=2$ and $k=4$. For $k=2$, all subjects expected a high probability of success in the first round, whereas a majority of the subjects expected

less than 60% chance for success when $k = 4$. While these beliefs cannot be directly validated, the results point to the importance of the required number of volunteers in shaping the subjects' behavior.

Moreover, the decay rate γ, which determines the weight accorded to the prior and previous observations compared to the most recent observation, was almost always above .95, with a mean of .93 and a median of .97 (Figure 5.6h). Only three subjects had a decay rate less than .95 (not shown in the figure), suggesting that almost all subjects relied on observations made across multiple rounds when computing their beliefs rather than reasoning based solely on the current or most recent observations.

5.6 Discussion

We introduced a normative model based on POMDPs for explaining human behavior in a group decision making task. Our model combines probabilistic reasoning about the group with long-term reward maximization by simulating the effect of each action on the future behaviour of the group. The greater accuracy of our model in explaining and predicting the subjects' behaviour compared to the other models suggest that humans make decisions in group settings by reasoning about the group as a whole. This mechanism is analogous to maintaining a theory of mind about another person, except the theory of mind pertains to a group member on average.

This is the first time, to our knowledge, that a normative model has been proposed for a group decision making task. Existing models to explain human behavior in the Public Goods Game, for example, are descriptive and do not provide insights into the computational mechanisms underlying the decisions [162]. While the regression-based descriptive method we compared our POMDP model to can potentially be seen as a "learned" model-free approach to mapping observations to choice in the next round, our model was able to outperform this method as well.

In addition to providing a better fit and prediction of the subject's behavior, our model, when fit to the subject's actions can predict success rate in each round without being explicitly trained for such predictions, in contrast to the other methods. Also, as alluded to in figures 5.6c, 5.6d and 5.6h, when fit to the subjects' actions, the parameters were all within a reasonable range, showing

the importance of prior knowledge and multiple observations in decision making. The POMDP model is normative and strictly constrained by probability theory and optimal control theory. The beta distribution is used because it is the conjugate prior of the binomial distribution [101] and not due to better fits compared to other distributions.

The POMDP policy aligns with our intuition about the action selection in the Volunteer's Dilemma task. A player chooses to free ride for two reasons: (i) when the cooperativeness of the group is low and therefore there is no benefit in contributing, and (ii) when the player knows there are already enough volunteers and contributing leads to a waste of resources. The two purple areas of Figure 5.4a represent these two conditions for $k = 2$. The upper left part represents large α_t and small β_t, implying a high contribution rate, while the bottom right part represents small α_t and large β_t implying a low contribution rate. When $k = 4$, all but one of the 5 players must contribute for group success - this causes a significant difference in the optimal POMDP policy compared to the $k = 2$ condition. As seen in Figure 5.4b, there is only a single region of belief space for which free-riding is the best strategy, namely, when the player does not expect contributions by enough players (relatively large β_t). On the other hand, as expected, this region is much larger compared to the same region for $k = 2$ (see Figure 5.4a). The POMDP model predicts that free-riding is not a viable action in the $k = 4$ case (Figure 5.4b) because not only does this action require all the other 4 players to contribute in order to generate the group reward in the current round, but such an action also increases the chances that the group contribution will be lower in the next round, resulting in lesser expected reward in future rounds. The opposite situation can also occur especially when $k = 2$. A player may contribute not to gain the group reward in the current round, but to encourage others to contribute in the next rounds. When an optimal player chooses free-riding due to low cooperativeness of the group, the estimated average contribution is so low that the group is not likely to get the group reward in the next rounds even with an increase in the average contribution due to the player's contribution. On the other hand, when an optimal player chooses to free-ride due to high cooperativeness of the group, the estimated average contribution rate is so high that the chance of success remains high in future rounds even with a decrease in average contribution rate due the player free-riding in the current round.

In a game with a predetermined and known number of rounds, even if the player considers the future, one might expect the most rewarding action in the last rounds to be free riding as there is little or no future to consider. However, our experimental data did not support this conclusion. Our model is able to explain this data using the hypothesis that subjects may employ a longer horizon than the exact number of rounds in a game. Such a strategy provides a significant computational benefit by making the policies for different rounds similar to each other, avoiding re-calculation of a policy for each single round. Recent studies in human decision making have demonstrated that humans may use such minimal modifications of model-based policies for efficiency [96, 131]. More broadly, group decision making occurs among groups of humans (and animals) that live together. Thus, any group decision making is practically infinite-horizon, i.e., there is always a future interaction even after the current task has ended, justifying the use of long horizons.

In the Volunteer's Dilemma, not only is the common goal not reached when there are not enough volunteers, but having more than the required number of volunteers leads to a waste of resources. As a result, an accurate prediction of others' intentions based on one's beliefs is crucial to make accurate decisions. This gives the model-based approach a huge advantage over model-free methods in terms of reward gathering, thus making it more beneficial for the brain to endure the extra cognitive cost. It is possible that in simpler tasks where the accurate prediction of minds is less crucial, the brain adopts a model-free approach.

Our model was based on the binomial and beta distributions for binary values due to nature of the task, but it can be easily extended to the more general case of a discrete set of actions using multinomial and Dirichlet distributions [3]. Additionally, the model can be extended to multivariate states, e.g., when the players are no longer anonymous. In such cases, the belief can be modeled as a joint probability distribution over all parameters of the state. This however incurs a significant computational cost. An interesting area for future research is investigating whether under some circumstances, humans model group members with similar behaviour as one subgroup in order to reduce the number of minds one should reason about.

Our POMDP framework assumes that each subject starts with the same prior about average group member contribution probability at the beginning of each game. However, subjects might try

to estimate this prior for a new group in the first few rounds, i.e., ”explore” their new environment, before seeking to maximize their reward (”exploit”) based on this prior [51]. Such an ”active inference” approach has been studied in two-person interactions [100, 99] and is an interesting direction of research in group decision making.

Mimicking human behavior does not guarantee that POMDP (or any model) is being implemented in the brain. However, POMDP's generalizability and the interpretability of its components such as existence of a prior or simulation of the future, make it a useful tool for understanding the decision making process.

The POMDP framework can model social tasks beyond economic decision making, such as prediction of others' intentions and actions in everyday situations [146]. In these cases, we would need to modify the model's definition of the state of other minds to include dimensions such as valence, competence, and social impact instead of propensity to contribute monetary units as in the PGG task [147].

The interpretability of the POMDP framework offers an opportunity to study the neurocognitive mechanisms of group decision making in healthy and diseased brains. POMDPs and similar Bayesian models have previously proved useful in understanding neural responses in sensory decision making [121, 68] and in tasks involving interactions with a single individual [166, 69, 5]. We believe the POMDP model we have proposed can likewise prove useful in interpreting neural responses and data from neuroimaging studies of group decision making tasks. Additionally, the model can be used for Bayesian theory-driven investigations in the field of computational psychiatry [70]. For example, theory of mind deficits are a key feature of autism spectrum disorder [7] but it is unclear what computational components are impaired and how they are affected. The POMDP model may provide a new avenue for computational studies of such neuropsychiatric disorders [136].

5.7 Methods

Experiment

30 right-handed students at the University of Parma were recruited for this study. One of them aborted the experiment due to anxiety. Data from the other 29 participants were collected, analyzed, and reported. Based on self-reported questionnaires, none of the participants had a history of neurological or psychiatric disorders. This study was approved by the Institutional Review Board of the local ethics committee from Parma University (IRB no. A13-37030), which was carried out according to the ethical standards of the 2013 Declaration of Helsinki. All participants gave their informed written consent. As mentioned in Results, each subject played 14 sessions of the Public Goods Game (PGG) (i.e., the Volunteer's Dilemma), each containing 15 rounds. In the first 2 sessions, subjects received no feedback about the result of each round. However, in the following 12 sessions, social and monetary feedback were provided to the subject. The feedback included the number of contributors and free riders, and the subject's reward in that round. Each individual player's action, however, remained unknown to the others. Therefore, individual players could not be tracked. We present analyses from the games with feedback.

In each round (see Figure 5.1), the participant had to make a decision within three seconds by pressing a key; otherwise the round was repeated. 2.5 to 4 seconds after the action selection, the outcome of the round was shown to the subject for 4 seconds. Then, players evaluated the outcome of the round before the next round started. Subjects were told that they were playing with 19 other participants located in other rooms. Overall, 20 players were playing the PGG in 4 different groups simultaneously. These groups were randomly chosen by a computer at the beginning of each session. In reality, subjects were playing with a computer. In other words, a computer algorithm was generating all the actions of others for each subject. Each subject got a final monetary reward equal to the result of one PGG randomly selected by the computer at the end of the study.

In a PGG with $N = 5$ players, we denote the action of player i in round t with the binary value of a_i^t $(1 \leq i \leq N)$ with $a_i^t = 1$ representing contribution and $a_i^t = 0$ representing free-riding.

The human subject is assumed to be player 1. We define the average contribution rate of others $\bar{a}^t_{2:N} = \frac{\sum_{i=2}^{N} a^t_i}{N-1}$ and generate each of the $N-1$ actions of others in round t using the following probabilistic function:

$$logit(\bar{a}^t_{2:N}) = e_0 a^{t-1}_1 + e_1((\frac{1-K^{T-t+1}}{1-K})^{e_2}\bar{a}^{t-1}_{2:N} - K). \tag{5.4}$$

where $K = k/N$ where k is the required number of contributors.

This model has 3 free parameters: e_0, e_1, e_2. These were obtained by fitting the above function to the actual actions of subjects in another PGG study [108], making this function a simulation of human behavior in the PGG task. Specifically, to generate the actions of others, we fixed e_2 to 1 for all games. e_0 was drawn randomly from the range of $[.15, .35]$ for each game and e_1 was set to $1-e_0$. This combination and the random sampling of e_0 in each game simulated different response strategies for the others in each game, simulating new sets of group members. Higher values of e_0 make the algorithm more likely to choose its next action based on the result of the group interaction in the previous round (especially the action of the subject). On the other hand, lower values of e_0 make the algorithm more likely to stick to its previous action. For the first round of each game, we used the mean contribution rate of each subject as their fellow members' decision.

POMDP for binary public goods game

The state of the environment is represented by the average cooperativeness of the group, or equivalently, the average probability θ of contribution by a group member. Since θ is not observable, the task is a POMDP and one must maintain a probability distribution (belief) over θ. The Beta distribution, represented by two free parameters (α and β), is the conjugate prior for binomial distribution [101]. Therefore, when performing Bayesian inference to obtain the belief state over θ, combining the Beta distribution as the prior belief and the binomial distribution as the likelihood results in another Beta distribution as the posterior belief. Using the Beta distribution for the belief state, our POMDP turns into an MDP with a two-dimensional state space represented by α and β. Starting from an initial belief state $Beta(\alpha_1, \beta_1)$ and with an additional free parameter γ, the next belief states is determined by the actions of all players at each round as described in Results.

For the reward function, we used the monetary reward function of the Public Goods Game (PGG). Therefore, the elements of our new MDP derived from the PGG POMDP are as following:

- $S = (\alpha, \beta)$
- $A = \{c, f\}$
- $T(s', s, a) : \begin{cases} P((\gamma\alpha + k' + 1, \gamma\beta + N - 1 - k')|(\alpha, \beta), c) = \binom{N-1}{k'} \frac{B(\gamma\alpha+k', \gamma\beta+N-1-k')}{B(\gamma\alpha, \gamma\beta)} \\ P((\gamma\alpha + k', \gamma\beta + N - k')|(\alpha, \beta), f) = \binom{N-1}{k'} \frac{B(\gamma\alpha+k', \gamma\beta+N-1-k')}{B(\gamma\alpha, \gamma\beta)} \end{cases}$
- $R(s, a) : \begin{cases} R((\alpha, \beta), c) = E - C + \sum_{k'=k-1}^{N} \binom{N-1}{k'} \frac{B(\alpha+k', \beta+N-1-k')}{B(\alpha, \beta)} G \\ R((\alpha, \beta), f) = E + \sum_{k'=k}^{N} \binom{N-1}{k'} \frac{B(\alpha+k', \beta+N-1-k')}{B(\alpha, \beta)} G \end{cases}$

$B(\alpha, \beta)$ is the normalizing constant: $B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$.

The POMDP model above assumes that the hidden state, i.e. θ, is a random variable following a Bernoulli distribution which changes with the actions of all players in each round. These actions serve as samples from this distribution, with α_1 and β_1 being the initial samples. Also, the decay rate γ controls the weights of previous samples. Using Maximum likelihood Estimation (MLE), for any t, θ_t equals $\alpha_t / (\alpha_t + \beta_t)$. One can also estimate θ in a recursive fashion:

$$\theta_{t+1} \leftarrow \frac{1}{\gamma\alpha_t + \gamma\beta_t + N}((\gamma\alpha_t + \gamma\beta_t)\theta_t + \sum_{i=1}^{N} a_i^t) \tag{5.5}$$

where a_i^t is the action of player i in round t ($a_i^t = 1$ for contribution and 0 for free-ride).

According to the experiment, the time horizon should be 15 time steps. However, we found that a longer horizon ($H = 50$) for all players provides a better fit to the subjects' data, potentially reflecting an intrinsic bias in humans for using longer horizons for social decision making. For each subject, we found α_1, β_1, and γ that made our POMDP's optimal policy fit the subject's actions as much as possible. For simplicity, we only considered integer values for states (integer α and β). The fitting process involved searching over integer values from 1 to 200 for α_1 and β_1 and values between 0 to 1 with a precision of .01 $(.01, .02, \ldots, .99, 1.0)$ for γ. The fitting criterion

was round-by-round accuracy. For consistency with the descriptive model, the first round was not included (despite the POMDP model's capability for predicting it). Since the utility value for public good for a subject can be higher than the monetary reward due to social or cultural reasons [44], we investigated the effect of higher values for the group reward G in the reward function of the POMDP. This however did not improve the fit.

As specified above, the best action for each state in round t is $U^t(s)$. The probability of contribution (choice probability) can be calculated using a logit function: $1/(1+exp(z(Q^t(s,f)-Q^t(s,c))$ [144]. For each k, we used one free parameter z across all subjects to maximize the likelihood of contribution probability given the experimental data (implementation by Scikit-learn [112]). Note that the parameter z does not affect the accuracy of fits and predictions since it does not affect the action with the maximum expected total reward.

In round t, if the POMDP model selects the action "contribution", the probability of success can be calculated as $\sum_{m=k-1}^{N-1} P(m|\alpha_t,\beta_t)$ (see Equation 5.3). Otherwise, the probability of success is $\sum_{m=k}^{N-1} P(m|\alpha_t,\beta_t)$. This probability value was compared to the actual success and failure of each round to compute the accuracy of success prediction by the POMDP model.

Model-free nethod: Q-Learning

We used Q-learning as our model-free approach. There are two Q values in the PGG task, one for each action, i.e., Q(c) and Q(f) for "contribute" and "free-ride" respectively. At the beginning of each PGG, Q(c) and Q(f) are initialized to the expected reward for a subject for that action based on a free parameter p which represents the prior probability of group success. As a result, we have:

$$\begin{cases} Q^1(c) \leftarrow p(E-C+G)+(1-p)(E-C) \\ Q^1(f) \leftarrow p(E+G)+(1-p)E \end{cases} \tag{5.6}$$

We customized the utility function for each subject by making the group reward G a free parameter to account for possible prosocial intent [44]. Moreover, as the probability of success is different for $k=2$ and $k=4$, we used two separate parameters p_2 and p_4 instead of p, depending on the value of k in the PGG.

In each round of the game, the action with the maximum Q value was chosen. The Q value for that action was then updated based on the subject's action and group success/failure, with a learning rate η^t. This learning rate was a function of the round number, i.e. $\eta^t = \frac{1}{\lambda_0 + \lambda_1 t}$ where λ_0 and λ_1 are free parameters and t is the number of the current round. Let the subject's action in round t be a^t, the Q-learning model's chosen action be $\hat{a}^t$, and the reward obtained be r^t. We have:

$$1 \leq t \leq 15 : \begin{cases} \hat{a}^t = \arg\max_{a \in \{c,f\}} Q^t(a) \\ Q^{t+1}(a^t) \leftarrow (1-\eta^t) Q^t(a^t) + \eta^t r^t \end{cases} \tag{5.7}$$

For each subject, we searched for the values of λ_0, λ_1, the group reward G, and the probability of group success p_2 or p_4 that maximize the round-by-round accuracy of the Q-learning model. Similar to the other models, the first round was not included in this fitting process.

Descriptive model

Our descriptive model was based on a logistic regression (implementation by Scikit-learn [112]) that predicts the subject's action in the current round based on their own previous action and the total number of contributions by the others in the previous round. As a result, this model has 3 free parameters (two features and a bias parameter). Let a_1^t be the subject's action in round t and $a_{2:N}^t$ be the actions of others in the same round. The subject's predicted action in the next round $t+1$ is then given by:

$$\hat{a}_1^{t+1} = \begin{cases} c & \kappa_0 + \kappa_1 a_1^t + \kappa_2 (\sum_{i=2}^N a_i^t) > 0 \\ f & \text{otherwise} \end{cases} \tag{5.8}$$

We used one separate regression model for each subject. As the model's predicted action is based on the previous round's actions, the subject's action in the first round cannot be predicted by this model.

Leave-one-out cross validation

For all three approaches, LOOCV was computed based on the games played by each subject. For each subject, we set aside one game, fitted the parameters to the other 11 games, and computed the

error of the model with fitted parameters on the game that was set aside. We repeated this for all games and reported the average of the 12 errors as LOOCV error for the subject.

Static probability distribution and greedy Strategy

If a player does not consider the future and solely maximizes the expected reward in the current round (greedy strategy) or ignores the effect of an action on others, the optimal action is always free-riding independent of the average probability of contribution by a group member. This is because free-riding always results in one unit more monetary reward (3 MU for success or 1 MU for failure) compared to contribution (2 MU or 0 MU), except in the case where the total number of contributions by others is *exactly* $k-1$. In the latter case, choosing contribution yields 1 unit more reward (2 MU) compared to free-riding (1 MU). This means that the expected reward for free-riding is always more than that for contribution unless the probability of observing exactly $k-1$ contributions by others is greater than .5. We show that this is impossible for any value of θ. First, note that the probability of exactly $k-1$ contributions from $N-1$ players is maximized when $\theta = (k-1)/(N-1)$. Next, for any θ, the probability of $k-1$ contributions from $N-1$ players is:

$$P(k-1|\theta) = \binom{N-1}{k-1}\theta^{k-1}(1-\theta)^{N-k} \leq \binom{N-1}{k-1}(\frac{k-1}{N-1})^{k-1}(\frac{N-k}{N-1})^{N-k} = .75^3 < .5. \quad (5.9)$$

for $N = 5$ and for either $k = 2$ or $k = 4$.

Chapter 6

BAYESIAN THEORY OF COLLECTIVE DECISION MAKING

6.1 Introduction

Collective decision making is critical for survival in animals that forage as a group [25]. Even though humans are not "hunter-gatherers" any more, collective decision making has remained a crucial element of modern human society, as exemplified by the practice of trial by jury [72, 109] and dealing with a global crisis such as a pandemic [153]. Group decision making can become extremely challenging when there is no communication between group members. Despite this challenge, humans usually manage to coordinate with each other and make decisions successfully after a very few interactions. This ability has not been achieved by Artificial Intelligence yet, despite of its extra ordinary progress in recent years [19]. In fact, mechanisms involved in group decision making is a topic of interest in many fields such as Psychology, Neuroscience, Economics, and Artificial Intelligence [54, 161, 66].

Conformity or aligning one's actions with other group members is a behavior that has been widely observed in group decision making by biologists and psychologists [23, 157, 77], for example, in developing social norms [139, 160]. In fact, even in competitive situations, humans may mimic their opponent's behavior unintentionally [103]. In a collective decision making task, by definition, at least some amount of cooperation between different group members is required for producing utility. Conformity provides a mechanism for cooperation. However, in many situations, there is also some amount of competition between group members, making them cooperate strategically. For example, different players might prefer different outcomes. In these cases, conformity might be too naive and additional processes, such as prediction of others' actions, are required to gain more utility.

This chapter is mainly outsourced from [82].

The ability to infer others' intentions from their behavior, known as theory of mind (ToM), is believed to play a key role during social interaction and decision making [8]. In the Bayesian analogy of theory of mind, i.e. Bayesian ToM, the brain reasons about others' intention based on 1) the observed behavior, 2) an internal mapping from intention to behavior, 3) a prior belief about the intention of others, and finally 4) inferring the intention through applying Bayes rule to the observation, mapping, and the prior (1 to 3) [5, 53]. Importantly, when a sequence of interactions are involved, ones' own actions could change others' intention and consequently their future actions. Therefore, there should be also a recipe to link the future state of mind of other group members and consequently their actions to one's own current action.

Recent studies have shown that humans are capable of inferring others' current and even future states of the mind during social interactions [148]. Here we present a normative framework based on Bayesian ToM, utility maximization, and conformity as the link between current and future states and action, that explains human behaviour in collective decision making in groups. While there exist some studies suggesting the connection of theory of mind and conformity in social decision making, e.g., the opportunity for reciprocity in multi-round games [43, 145], there is no mathematical framework or quantitative analysis demonstrating this connection.

First, we present a Bayesian model of conformity as the basis for our framework for collective decision making. Then, we show how a (meta-)Bayesian agent can make better decisions (in terms of total utility gain) in complicated tasks with the presence of competitiveness between group members by reasoning about the belief of other Bayesian agents that utilize conformity. In addition, we show this framework can be extended to model recursive social reasoning and different "levels of theory of mind" in collective decision making. Following the terms used in the literature on two-person interactions [165, 36], a Bayesian agent in our framework that utilizes conformity is called level-0 ToM agent, a (meta-)Bayesian agent that reasons about other Bayesian agents that utilize conformity is level-1 ToM agent, a (meta-)Bayesian agent that reasons about other (meta-)Bayesian agents that reason about Bayesian agents that utilize conformity is level-2 ToM agent, and so on. This framework is a generalization of our previous level-1 ToM Bayesian model that explained human behavior in public good game in the previous chapter to multiple levels of theory

of mind and consequently, a broader range of collective decision making tasks. To our knowledge, this is the first mathematical connection between Bayesian ToM and conformity.

We tested our framework on three different collective decision making experiments involving human subjects: a consensus task, a thresholded public good game, and a prisoner's dilemma experiment, each conducted in two different conditions. Our normative Bayesian framework provided quantitative fits of human behavior, outperforming other models, on all of these tasks. Moreover, the levels of theory of mind that the subjects utilized in the experiments were aligned with their gained reward compared to other subjects. In addition, comparing different tasks with each other, overall fitted levels of ToM in each task were compatible with the components and nature of each task.

6.2 Theoretical results

We investigate the problem of collective decision making with $N > 2$ players and multiple rounds. In each round, the players choose their actions simultaneously and then, all actions in that round are shown to all players. The actions could be anonymous. The same set of actions is available to each player. In each round, each player chooses one action for all of the group, e.g. there is no individual punishment. More importantly, the reward (utility) of each player in each round depends only on their own action and how many of each of the available actions was selected by others in that round. For simplicity, we assume that the number of possible actions is 2, i.e., the set of actions $A = \{a1, a2\}$.

6.2.1 Bayesian conformity: matching the group

Conformity is matching the behavior of the whole group. As players might choose different actions and also due to stochastic nature of human behavior, we model the behavior of the group in a probabilistic fashion: An average group member's action follows a Bernoulli distribution with θ as the parameter/mean of choosing action $a1$. In other words, an average group member chooses action $a1$ with probability of θ. As a result, a player using conformity as their strategy should

also choose $a1$ with probability θ. In the Bayesian ToM language, θ represents the intention of choosing $a1$ by an average group member (state of the "mind of the group") and the Bernoulli distribution is the mapping from intention to action. Consequently, the likelihood of observing m $a1$ actions in total from N players is given by the Bernoulli probability density function, i.e:

$$P(m|\theta) = \binom{N}{m} \theta^m (1-\theta)^{N-m}. \tag{6.1}$$

The rational behind the concept of average group member is that tracking individuals are not useful, or sometimes not even possible due to anonymity. All that matters is the group as a whole.

The state parameter, θ, is not observable to the players. Each round only provides indirect information about this latent variable via the actions of all N players. As a result, we assume that the player maintains a "belief" about θ, i.e., they maintain a probability distribution over $\theta, P(\theta)$. Starting from a prior belief, which could be shaped from previous life experience or fictitious play, the player updates their belief about θ after each round of the game based on $N-1$ actions of others as well as their own action (the player themselves is a member of the group) via Bayes rule.

Because θ represents a Bernoulli distribution, we express the belief of the player about θ with a Beta distribution, which is determined by two parameters α and β:

$$\begin{aligned} &Beta(\alpha,\beta) : P(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta \end{aligned} \tag{6.2}$$

This choice of representation makes the belief of the player remain Beta distribution(Beta distribution is the *conjugate prior* for Binomial distribution). The posterior probability of θ after observing m actions of $a1$ from all N players is $Beta(\alpha+m, \beta+N-m)$ [102]. Note that in this context α represents total previous samples of $a1$ and β is the total number of $a2$ samples. Consequently, α to β ratio represents the proportion of samples, or how much $a1$ is more (or less) favorable than $a2$. Moreover, generally larger α and β means new observations (samples) change their ratio less.

As the task has multiple rounds, the player starts with a prior probability of $Beta(\alpha_1, \beta_1)$, updates it after each round, and uses the posterior probability of θ as the prior for the next round.

This is a Hidden Markov Model (HMM) where the player infers the state of mind of the group about the next action by observing the previous actions [102].

Due to possible changes in other group members' strategies, the most recent observations are often more reliable, thus deserving a larger weight in the inference. We model this by using a decay rate $0 \leq \lambda \leq 1$ for the prior. This means that a prior of $Beta(\alpha, \beta)$ with m out of N actions being $a1$ in the current round results in a posterior of $Beta(\lambda\alpha + m, \lambda\beta + N - m)$. For example, $\lambda = 0$ means that only the most recent round determines the posterior and $\lambda = 1$ considers all previous rounds equally important.

At the beginning of each round, the player chooses an action. According to the principle of conformity, $a1$ should be chosen with probability of θ. As the player has a posterior probability over θ instead of the exact value, they use the expected value of θ, which is $\alpha/(\alpha+\beta)$ [102]. We model this scenario with an HMM (instead of a decision process - see below) as the player does not consider the effect of their own action on others and the reward function. Figure 6.1a shows the graphical model of this process, where m_t represents the total number of action $a1$ in round t.

In summary, a player that utilizes Bayesian conformity has a prior belief of $Beta(\alpha_1, \beta_1)$ over θ before the start of a multi-round game, with a decay rate λ. At round $t \geq 1$, the player chooses $a1$ with probability of $\alpha_t/(\alpha_t + \beta_t)$. Then, after observing everyone's actions in that round, if there are m_t $a1$ actions in total, the belief changes to $Beta(\alpha_{t+1}, \beta_{t+1})$ where $\alpha_{t+1} = \lambda\alpha_t + m_t$ and $\beta_{t+1} = \lambda\beta_t + N - m_t$.

6.2.2 Meta-Bayesian Conformity: Influencing the Group

The model in the previous section ignored the fact that the player's actions can potentially influence the actions of the group on average. If the group members also utilize conformity, one might be able to influence their future actions particularly if α and β are small, the decay rate is large, the values of α and β are close to each other (i.e., θ is around .5). As a result, a player who takes advantage of this knowledge can increase their total expected reward over the course of the game by leading the group to states that are more rewarding in the later rounds of the game. The idea of selecting actions that maximize the total expected reward transforms the model from one based on

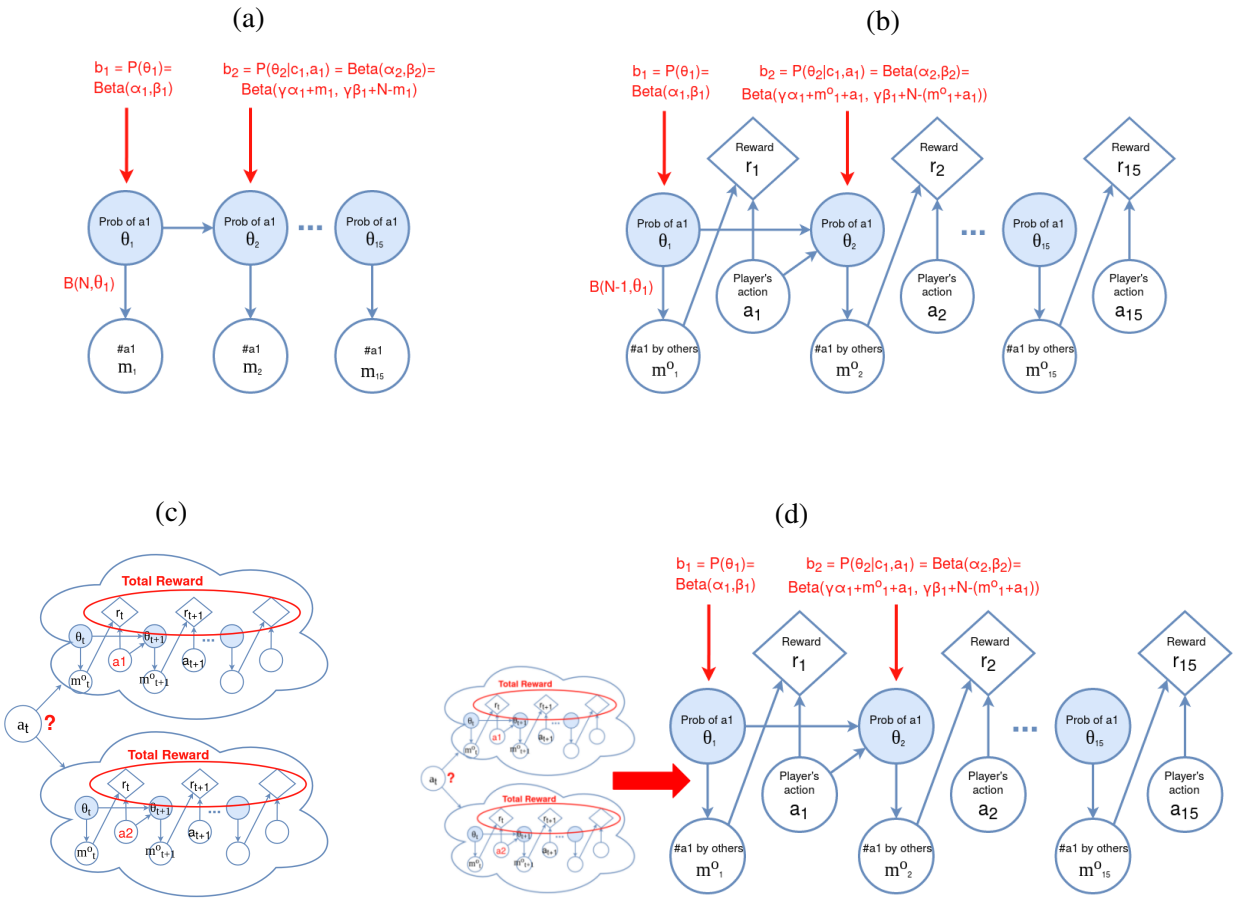


Figure 6.1: **Graphical models of different levels of ToM in collective group decision making**. **a)** A conformist, or level-0 ToM agent can be modeled as a Hidden Markov Model (HMM) where the decision making only depends on inference. b) Level-1 ToM agent that uses the conformity in the group to influence it and maximize the utility can be modeled with Partially Observable Markov Decision Process (POMDP). **c)** In order to maximize the reward, level-1 agent should simulate the future events based on each possible choice. **d)** Higher than 1 level of ToM can be also modeled by a POMDP in which the actions of others are derived from policy of the lower-level POMDP.

HMMs (previous section) to a Partially Observable Markov Decision Process (POMDP) [75].

Figure 2a shows the graphical model of the POMDP, which is basically the HMM with action of the player separated from others (m_t^o represents the total number of $a1$ by others in round t) and the reward function. Note that the reward depends solely on the action of the player and the total number of $a1$ actions by others. In each step, the agent that utilizes the POMDP model, i.e. the "POMDP solver", chooses an action that maximizes its expected total reward in the future. This requires *mental simulation* of future events after choosing each action, similar to *model-based* planning in reinforcement learning literature [61] (Figure 6.1c).

The key element of this process is how each action changes the state of the task(here the intention of the group). If the group utilize conformity, each action by the player increases the intention of choosing that action in the next round by others. Let the player's belief be $Beta(\alpha,\beta)$. If m other players choose $a1$, the belief of the player in the next round would be $Beta(\gamma\alpha+m+1,\gamma\beta+N-1-m)$ if their own action is also $a1$. However, the belief would be $Beta(\gamma\alpha+m,\gamma\beta+N-m)$ if the player's own action is $a2$. As mentioned before, this could make a notable difference in the long run, especially when α and β are small or close to each other. Using the mental simulation based on the transition function, a player that utilizes POMDP comes up with an optimal *policy*, $\pi_t^*(\alpha_t,\beta_t)$, that determines the action for each belief state and round. In fact this POMDP, is the same as our model in the previous chapter where $a1$ is contribution in that model and $a2$ is free-ride.

6.2.3 Higher levels of theory of mind

A Bayesian agent that utilizes conformity only infers the state of mind of others without assuming others have the same inference capability as well (level-0). The POMDP described above assumes others infer the state of mind of group members as well (level-1) to the extent of matching their probability of actions. We extend this reasoning here to achieve higher levels of theory of mind. A level-k agent is a POMDP agent that assumes others are level-$(k-1)$ agent. The Interactive-POMDP (I-POMDP) model is a general framework for modeling other POMDP agents with arbitrary transition, observation, and reward functions [55]. If the rules of the task are conveyed to all

players, achieving higher levels of ToM becomes more computationally tractable as the agent uses the same transition and observation functions for all members. A significant practical problem, however, is that the reward function of others is not often known. As a result, modeling higher levels of ToM becomes more plausible when the reward function is (at least mostly) similar for all group members, e.g., in the case of monetary rewards.

If the player uses a common reward function for all members of the group, the level-k agent ($k > 1$) is modeled as a POMDP very similar to a level-1 POMDP, but with a different assumption about others' actions. Instead of conformity, level-k agent assumes that others act according to the policy of level-$(k-1)$ ToM POMDP of the same model from now on. For example, with the belief state of (α_t, β_t) level-2 ToM agent assumes other players act according to $\pi_t^*(\alpha_t, \beta_t)$. This means that to calculate level-k policy, one should calculate level-$(k-1)$ policy, and consequently all the way down to level-1 policy.

If the player does not know the reward function of others, they could estimate it based on the dynamics of the game. When a player's level of reasoning is higher than 0, they know their actions could lead the group towards selecting actions that produce more reward for themselves. As a result, when a level-1 or higher level player chooses an action, either the immediate reward for that action is higher for them, or due to the state of the game, that action produces more expected reward despite producing less immediate reward. In the latter case, the chosen action would not change if one assumes a higher reward for it. As a result, for level-k agent ($k > 1$), when the belief state is (α_t, β_t), the player can divide other players into two groups based on their "preference" (immediate reward) for an action and estimate the reward function of each group separately. More specially, $\alpha_t/(\alpha_t, \beta_t)$ of other players prefer action $a1$ and the rest prefers $a2$. Similar to the previous case, level-k player assumes that both groups utilize level-$(k-1)$ POMDP. Formal definition of both of these POMDPs are explained in detail in Methods.

6.3 Experimental results

We tested our framework on the human behavioral data from three different collective decision making experiments. The first was a consensus group decision making task [144] where $N = 6$ or

$N = 4$ players need to agree on one of the two items presented to them within a limited number of rounds. The second experiment was the volunteer's dilemma experiment explained in the previous chapter. The last experiment was a Prisoner Dilemma (PD) task in a network where each subject played the game with 4 other players for 50 rounds [59]. This experiment were held in two different sessions with different types of information provided to players.

We fit models based on conformity with different levels of ToM to the behavior of each subject and compared the accuracy of the different level models in explaining human behavior. The accuracy of a model was determined by the similarity between the model's predicted action and the actual action of the subject in each round on average. In other words, if the predicted action of a model was $\hat{a}$ and the real action was a in a round, the average error was the average of the binary error $|\hat{a} - a|$ over all rounds of all games for the subject (accuracy = 1 - average error). In the level-0 agent, similar to classification methods, and to produce comparable results for higher levels, the selected action was $a1$ when $\alpha/(\alpha+\beta)$ was more than .5, and $a2$ otherwise.

In addition to fitting accuracy, in the first two experiments where the subjects played more than one game, we calculated Leave-One-Out Cross Validation (LOOCV) accuracy where at each iteration, the left-out data point was one whole game. LOOCV calculation was not possible for the third task, as each player only played one long game. We compared both fitting and LOOCV accuracy of a reinforcement-based model-free approach to our framework [93, 97]. In this approach, the player chooses the most rewarding action according to rewards in previous rounds: the agent starts the task with an initial value for each action, chooses the action with the maximum value in each round, and updates its value based on the gained reward in that round with a weight called the learning rate [152].

6.3.1 Consensus decision making

Our first analysis is on the behavioral data from the consensus decision making experiment conducted by [144]. In this experiment, each of 120 subjects played the game 40 times, 20 with 5 other players and 20 with 3 other players. Each game started with the presentation of two options to the players. The subjects had to choose one of them in each round. The game ended when

all players chose the same option, otherwise it went to the next round with the same two options. After each round, each player observed the others' selected actions as red dots under each option, meaning that individual actions were anonymous. After the 10th round, if consensus had still not been reached, the game ended with a probability of 25% in each subsequent round. If the players reached a unanimous consensus, they all received the chosen option. If the game ended without a consensus, subjects did not receive anything. Before the experiment, subjects' preference or value for each item (between \$0 to \$4) was determined through a Becker-DeGroot-Marschak (BDM) auction [11]. More details of the task can be found in the original article describing this experiment [144].

We analyzed the games that lasted more than 1 round for each subject. As each player did not know the values of other group members for each item, we used the actions from the first round as the prior for the rest of the game, i.e., $\alpha_1 = m_0$ and $\beta_1 = N - m_0$ (equivalent to considering no prior knowledge at the beginning) where m_0 is the number of players that chose option 1 in the first round for all levels of ToM.

For level 1 and higher the value of options should be considered in the model as well. The reward for each option for each player was set to the value of that item that was determined in the auction process before the experiment. For the level-2 and higher ToM models, as the subjects did not know the others' values for each item, we used two reward functions, representing a preference to each choice (see Methods). We assumed that the subject estimated the value of each action for others as \$2.5 for their favorite option and \$1.5 for the other option (the median value of \$2 $\pm$\$.5). As a result, for all levels of our model, there was only one free parameter, the decay rate λ, for each subject. More details of model fitting are presented in Methods.

As shown in Figure 6.2a, the behavior of 82 of the 120 subjects in the experiment were better or equally well explained by the level-0 model compared to the level-1 model in terms of higher fitting accuracy. Average fitting accuracy was 78.0% ($SD = 0.093$) and 72.3% ($SD = .126$) for level 0 and level-1 models respectively. Average LOOCV accuracy was 75.0% ($SD = 0.102$) for level 0, and 70.5% ($SD = .125$) for level 1. Importantly, for almost all subjects with higher fitting accuracy of level-1 compared to level-0 (almost all green points in figure 6.2b are above dotted

$x = y$ line), LOOCV accuracy of level-1 was also higher than LOOCV accuracy of level-0 and vice versa (almost all purple points in figure 6.2b are below dotted $x = y$ line).

A model with a mix of multiple levels was constructed by choosing the level with higher LOOCV accuracy for each subject, since it is a more reliable measurement compared to fitting accuracy. The tie was broken first in favor of the level with lower fitting accuracy (since the drop in accuracy was lower) and then in favor of the lower level. The average accuracy of mixed model of level 0 and 1 was 79.1% accuracy ($SD = 0.092$) and its average LOOCV accuracy was 77.3% ($SD = 0.096$).

The level-2 model had a lower fitting accuracy ($mean = 73.4\%$, $SD = 0.108$) compared to the level-0 model in almost all subjects (Figure 6.2c). Its LOOCV accuracy for several subjects, however, was a little higher than the mix of level 0 and 1 ($mean = 71.7\%$, $SD = 0.110$, Figure 6.2d). Since, this improvement was very small, adding level 2 to the mixed model, did not change overall fitting accuracy (given our measurement precision). Moreover, it improved the average LOOCV accuracy only 0.4%. Finally, both fitting and LOOCV accuracy of the level-3 model is lower than the mix of level 0 and 1, for almost all subjects (Figures 6.2e and 6.2f). Overall, it is more reasonable to assume subjects utilize only level 0 and 1 ToM.

We also fit Q-learning [71], a model-free reinforcement learning model with three free parameters to the subjects' behavior. The first free parameter was a reward for reaching a consensus, which was added to the utility of choices in the auction. The other two free parameters determined the learning rate through time (more details in Methods).

The average fitting accuracy of this model was 61.9% ($SD = 0.131$), significantly worse than the fitting accuracy of both level-0 and level-1 ToM models in our framework (Wilcoxon signed-rank test, compared to level-0: $p = 3.1 \times 10^{-16}$, compared to level-1: $p = 2.3 \times 10^{-11}$), and thus also significantly worse than the mix of these two ToM models (Wilcoxon signed-rank test, $p = 4.5 \times 10^{-18}$). In addition, the average LOOCV accuracy of the model-free reinforcement learning model was 49.7% ($SD = 0.128$), significantly worse than LOOCV accuracy of both the level-0 and level-1 ToM models (Wilcoxon signed-rank test, compared to level-0: 7.1×10^{-20}, compared to level-1: 6.0×10^{-20}), and the mix of the two models: $p = 9.9 \times 10^{-21}$).

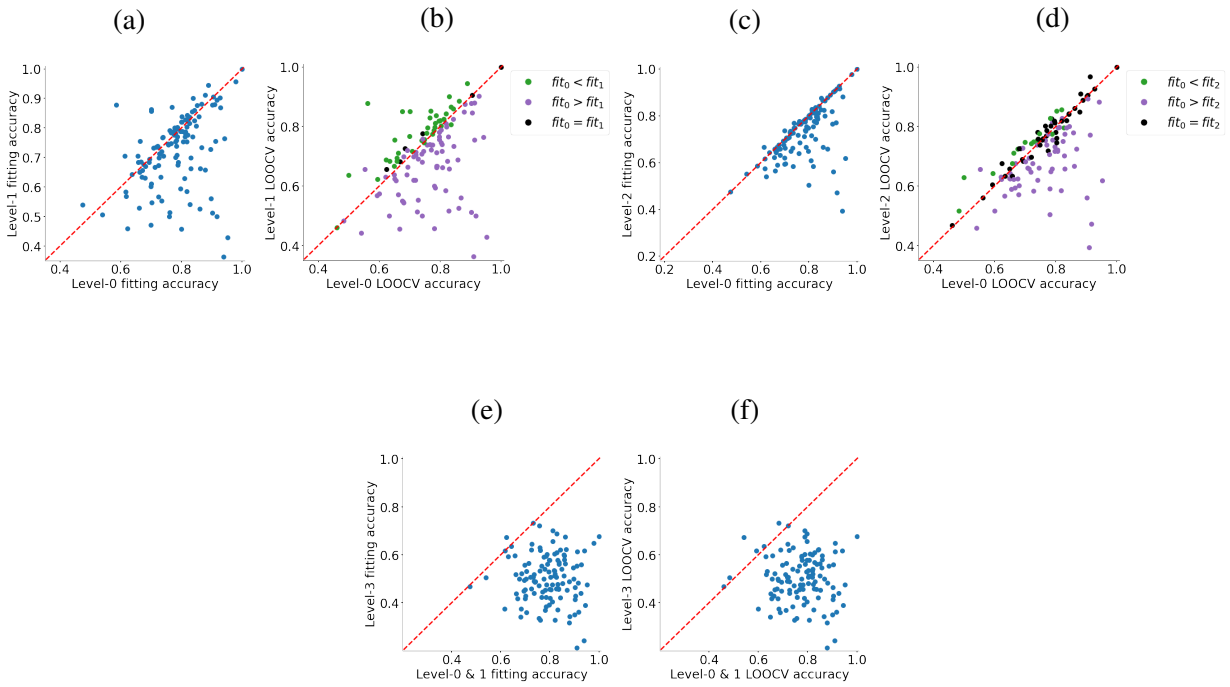


Figure 6.2: **Different levels of ToM in consensus decision making**. **a)** Comparison of level-0 and level-1 ToM fitting accuracy for each subject. The level 0 model explained a greater proportion of the subjects' behavior (see text for details). **b)** Comparison of level-0 and level-1 ToM LOOCV accuracy for each subject. Color each data point shows which level provides a better fitting accuracy for that point (green for 1, purple for level 0, and black for equal fit). **c)** Comparison of level-0 and level-2 ToM fitting accuracy for each subject. Level-0 explained the subjects' behavior better for almost of the subjects. **d)** Same as (c) but for level-2 versus mix of level 0 and 1. **e)** Comparison of level-3 and mix of level 0 and 1 ToM fitting accuracy for each subject. Mix of level 0 and 1 explained the subjects' behavior better for almost all of the subjects. **f)** Same as (e) but for LOOCV accuracy.

Comparing the choices of subjects for whom level-1 ToM provided the best fit with those for whom level-0 ToM explained their behavior better is also insightful. Subjects with behavior better

explained by level-1 ToM obtained their preferred choice (strictly greater value according to the auction) 52.2% ($SD = .138$) of the time on average, significantly higher than those whose behavior was better explained by level-0 ToM with 45.3% ($SD = .150$) (Mann-Whitney U two-sided test, $p = .011$). These results held when including level-2 in the analysis (Mann-Whitney U two-sided test, $p = .027$). Moreover, there was no significant difference between the average value difference of presented choices among subjects with different best fitted level of ToM (Mann-Whitney U two-sided test, $p = .84$). This shows that the difference between value of presented choices was not the reason behind obtained levels through fits.

6.3.2 Public goods game

Our second data set is from the experiment of Volunteer's Dilemma (VD) explained in the previous chapter. In this experiment, the reward function for all players is the same because monetary reward was used (rather than items of different desirability). As a result, players might use a prior based on their previous experience in life or fictional play, even before the start of the game. Since the required number of volunteers have a huge effect on the probability of success and consequently the need for approximating it, we used different set of parameters for each condition. As a result, for the models of all levels of ToM, there are 3 free parameters for each k and each subject in total, i.e., λ, α_1 and β_1.

As seen in Figures 6.3a, the level-1 model's fitting accuracy was higher than the level-0 for almost all subjects when $k = 2$ (Blue data points; level-0 fitting accuracy: $mean = 74.0\%$, $SD = 0.103$, level-1 fitting accuracy: $mean = 82.1\%$, $SD = 0.078$). On the other hand, level-0 model explained the behavior of many subjects better than level-1 model when $k = 4$ volunteers were needed (Orange data points; level-0 fitting accuracy: $mean = 86.0\%$, $SD = 0.104$, level-1 fitting accuracy: $mean = 85.4\%$, $SD = 0.085$). LOOCV accuracy followed the same trend with an overall shift in favor of level-0 model (Figure 6.3b; $k = 2$: level-0 LOOCV: $mean = 71.1\%$, $SD = 0.120$ and level-1 LOOCV: $mean = 72.4\%$, $SD = 0.116$; $k = 4$: level-0 LOOCV: $mean = 83.1\%$, $SD = 0.124$ and level-1 LOOCV: $mean = 78.2\%$, $SD = 0.129$)

Similar to the previous task, we built the mixed model of level 0 and 1 based on their best

LOOCV accuracy. Level-2 model had higher fitting accuracy compared to this mixed model for some subjects in both conditions. (figure 6.3c; level-2 fitting accuracy for $k = 2$: $mean = 80.9\%$, $SD = 0.083$ and for $k = 4$: $mean = 74.5\%$, $SD = 0.137$). This was also true for LOOCV as shown in figure 6.3d (level-2 LOOCV for $k = 2$: $mean = 70.6\%$, $SD = 0.099$ and for $k = 4$: $mean = 62.5\%$, $SD = 0.200$)

We built the mixed-model of levels 0 to 2 based on LOOCV and compared level-3 model with it. While level-3 produced a better fit for some subjects (figure 6.3e), its LOOCV accuracy was lower than the mixed-model for almost all subjects and conditions (figure 6.3f). As a result, we only used the first three levels for our further analysis. The average fitting accuracy of this mixed model was 82.2% ($SD = .075$) when $k = 2$ and 87.1% ($SD = .097$) when $k = 4$. Moreover, it had average LOOCV accuracy of 77.3% ($SD = .090$) when $k = 2$ and 84.7% ($SD = .109$) when $k = 4$.

We compared the model-free reinforcement learning (RL) model to the mixed model of level-0, 1 and 2 model. The RL model had 4 parameters in total for each k. The first parameter was a reward for generating public good, which was added to the monetary reward. The next parameter determined the chance of producing public good and was used to define the initial Q-value of each action. The final two free parameters determined the learning rate, similar to the consensus task (more details in the methods). The average fitting accuracy for the RL model was 73.0% ($SD = .103$) for $k = 2$ and 78.5% ($SD = .105$) for $k = 4$ significantly worse than the mixed model's fitting accuracies (Wilcoxon signed-rank test, $k = 2 : p = 1.8 \times 10^{-5}$, $k = 4 : p = 4.3 \times 10^{-5}$). Also, the average LOOCV accuracy of our mixed model was significantly higher than average LOOCV accuracy of the RL model which was 73.4% ($SD = .142$) when $k = 4$ (Wilcoxon signed-rank test, $p = 2.1 \times 10^{-5}$). When $k = 2$, the LOOCV accuracy of RL model with the average value of 75.1% ($SD = .109$) was not significantly different from our mixed model (Wilcoxon signed-rank test, $p = 0.09$)

There was a significant difference in contribution rate of subjects for whom the best fit was level-0 model compared to those for whom level 1 or 2 described their behavior better when $k = 4$ volunteers were needed. Contribution rate (per round) of subjects whom level-0 provided the best fit was on average .618 ($SD = .291$) when $k = 4$, significantly larger than contribution rate for

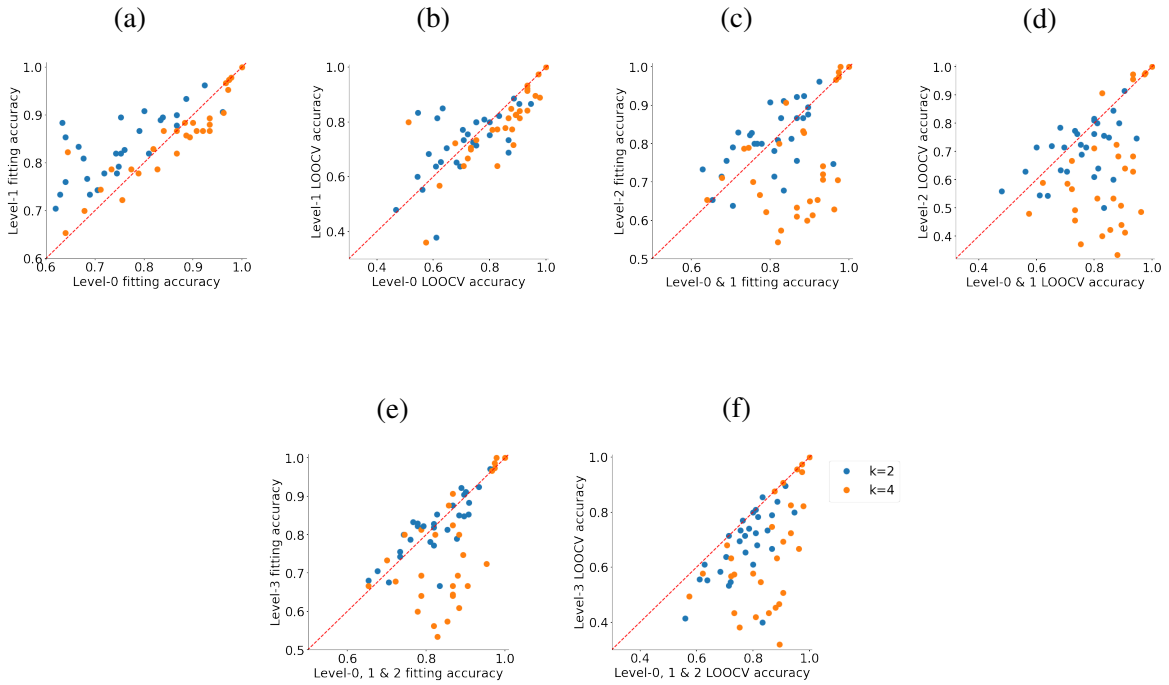


Figure 6.3: **Different levels of ToM in the volunteer's dilemma task. a)** Comparison of level-0 model and level-1 ToM model fitting accuracy for each subject. **b)** Comparison of level-0 model and level-1 ToM model LOOCV accuracy for each subject. **c)** Comparison of mixed model of level-0 and level-1 with level-2 ToM fitting accuracy for each subject. **d)** Comparison of mixed model of level-0 and level-1 with level-2 ToM LOOCV accuracy for each subject.**c)** Comparison of mixed model of level-0, level-1, and level-2 with level-3 ToM fitting accuracy for each subject.**c)** Comparison of mixed model of level-0, level-1, and level-2 with level-3 ToM LOOCV accuracy for each subject. Blue represents games where $k = 2$ and orange represents $k = 4$.

subjects with level 1 or 2 as their best fit with average of .227 ($SD = .165$) (Mann-Whitney U test two-sided test, $p = 0.011$). When $k = 2$, however, the contribution rate in different levels were not different.

6.3.3 Prisoner's dilemma

We analyzed the behavior of 144 subjects playing prisoner's dilemma with 4 other players in one long game with 50 rounds [59]. In each round of the game, the players chose between two colors, blue and yellow. Depending on the chosen color, the player received a different reward in the interaction with each of their group members, called their neighbors. Blue would lead to 5 units of reward if the neighbor's choice was also blue, and 0 otherwise (analogous to cooperation in the classical prisoner's dilemma). Yellow gave the player 6 units of reward if the neighbor's choice was blue, and 1 otherwise (analogous to defection). The total gained reward of each player was the sum of reward gained through their interactions with each of their neighbors.

The experiment consisted of two sessions. In one session with 80 subjects, players observed the action and gained reward of their neighbors after each round. In the second session with the other 64 subjects, only the action of players was visible to their group members. Importantly, each session was played in multiple 4 by 4 connected lattices with Von Neumann neighborhood [59]. This means that while each player played with the same set of people in each round, their neighbors did not share any neighbor with them. Moreover, each player's actions indirectly affected all other 15 players in that lattice. More details about the experiment can be found in the original study [59].

As each subject played only one long game, we could not perform any cross validation test. However, multiple aspects of this experiment make it a suitable test for evaluating our model. The structure of the lattice, and the visibility of each individual's actions (in both sessions) are different from our previous tasks and to some extend in contrast with some of the assumptions of our model. More importantly, the difference of the two sessions is the information given to each player, and thus more fundamental than the difference in the conditions of previous experiments (i.e. number of players or required contributors).

Similar to the public good game experiment, the reward function for all players is the same. Therefore, we considered free parameters of α_1 and β_1 for the prior, in addition to the decay rate γ for all of our levels. Subjects' behavior in this task is explained by higher range of levels of ToM compared to our previous experiments. As shown in Figure 2.2a, levels 1, 2 and maybe 3 of

our framework could explain different subject's behavior in both sessions. As each subject only participated in one of the sessions, we show subjects of each session with different color, i.e. blue for the players that only saw actions of others, and orange for those who had both utility and action of others after each round. Also, as our model's accuracy was the same for some players in both comparing levels, the size of data points grows proportional to the number of subjects with that fitting accuracy values in the panels.

Level-0 model produced average fitting accuracy of 77.8% ($SD = .127$) when only actions of others were shown and 78.2% ($SD = .119$) when both action and reward of each player were visible. Its accuracy was higher than level-1 for only 5 subjects (average fitting accuracy of level-1: 83.3% ($SD = .092$)) when only actions of others were shown (Blue data points in figure 6.4a). In the session where both action and reward of each player were visible, fitting accuracy of level-0 was higher than level-1 accuracy for only 11 subjects (Orange data points in figure 6.4a; average fitting accuracy of level-1: 83.2% ($SD = .093$)). Level-2 model had average fitting accuracy of 81.8% ($SD = .104$) when only actions were visible to others and 83.3% ($SD = 0.099$) when reward of each player was also shown to others. As shown in figure 6.4b, none of the level-1 and level-2 models outperformed the other one in terms of generating higher fitting accuracy for more number of subjects in any of the sessions.

Level-3 model produced average accuracies of 83.6% ($SD = .104$) and 81.3% ($SD = .112$) for sessions with and without reward of others as a part of available information, respectively. This model, provided a better fit for 47 subjects combined compared to the mixed model of level 1 and 2 (figure 6.4c). In many of these 47, however, level-3 model explained only one more round better (data points with center between red and green dotted lines). Also, as shown in figure 6.4d, except for 2 subjects, level-4 (average accuracies of 83.6% ($SD = .111$) with and 80.6% ($SD = .132$) without the reward of others as available information) model explained only one more round better when it outperformed the mixed of its previous levels. These results suggest that players mostly utilized level-1 or level-2 ToM in this experiment with the possibility of some utilizing level-3.

We compared the mixed model of level 1 and level 2 with a model-free RL method with four free parameters. The first parameter was a weight that each player assigned to others' reward. Total

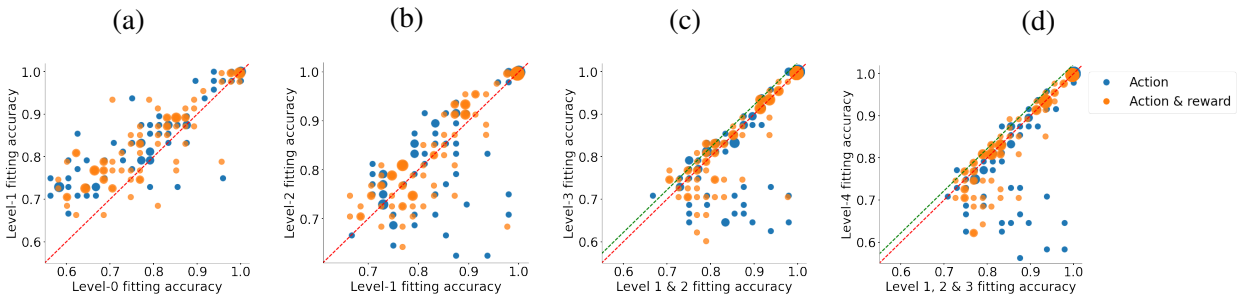


Figure 6.4: **Different levels of ToM in the prisoner's dilemma task. a)** Comparison of level-0 model and level-1 ToM model fitting accuracy for each subject. The level-1 model explained the behavior better for almost all subjects. **b)** Comparison of level-1 and level-2 models' fitting accuracy for each subject. Behavior of many subjects were better explained by level-1, and many by level-2 model **c)** Comparison of level-3 model with the mix (best) of level-1 and level-2 fitting accuracy. The level-3 model could explain the behavior of notable amount of players (42) better than the mix of level-1 and level-2. The fit of level-3 for most of these 42 players, however, was only 1 more round better ($1/49 \approx 2.04\%$). **d)** Comparison of level-4 ToM with level-1 to 3 ToM models. Level 4 explains the behavior of only 2 subjects better than 1 more round compare to level-1 to 3 ToM models. Blue data points represent the session that only actions of others were available to the player. Orange data points represent the session that gained rewards of others were also available. Red dotted lines are $y = x$ and green dotted line is $y = x + 0.022$. The size of data points grows proportional to the number of subjects with that accuracy values.

reward of others times this weight was added to the player's own reward. This parameter modeled empathy. The other parameter was the proportion of cooperative players (selecting blue) and were used to define the initial Q-value of each action. The last two determined learning rate similar to the RL approaches for other tasks (more details in methods). The mixed-model of level 1 and 2 with average fitting accuracy of 85.4% ($SD = 0.088$) for both sessions combined was significantly larger than the accuracy of the RL model with average of 79.0% ($SD = 0.117$) (Mann Whitney U

two-sided test, $p = 1.2 \times 10^{-6}$).

Interestingly, there is a meaningful relationship between subjects' fitted level of ToM and their amount of gained reward in the two sessions. Subjects with higher than 1 level of ToM received significantly larger reward in the session where the rewards of others were not shown compared to those with the same level of ToM in the other session. However, there was no difference between the gained reward of subjects with lower than level 2 ToM between the two sessions. When only actions were visible, subjects had to rely on their own reasoning to gain higher reward. As a result, they did not change their policy, which was totally hidden from others, through the session (Note that change of policy is different from change of action). Therefore, the assumptions of higher than 1 level players about others' hidden policy remained true. In the session that reward of each player was also conveyed to others, however, players could, and as the original study of this experiment showed did, imitate the more rewarding actions [59]. This imitation was a recipe of action selection different from their own strategy without such information (with the exception of most successfully player(s) in each group whom being imitated). This led to significant drop in the gained reward of players who planned based on others' policy (level-2 and higher). Note that each strategy in our framework is optimal if the underlying assumptions are true.

Level-0 and 1 ToM players do not plan on others' hidden policy. They plan based on others' belief obtained from the history of observable actions. As a result, lower than level-2 ToM players of both sessions, gained the same amount of reward on average. Figure 6.5a shows this phenomenon when assumed ToM levels are only 1 and 2 (Mann-Whitney U one-sided test; level 1: $p = 0.35$, level 2: $p = 0.0006$). Figure 6.5b shows it when level-3 ToM is also considered (Mann-Whitney U one-sided test; level 1: $p = 0.50$, level 2: $p = 0.007$, level 3: $p = .01$). In fact, it exists even if we consider all levels of 0 to 4, as shown in figure 6.5c (level 0: $p = 0.33$, level 1: $p = 0.45$, level 2: $p = 0.033$, level 3: $p = 0.035$, level 4: $p = 0.005$).

6.4 Simulation results

Higher levels of ToM in our framework assume deeper levels of optimality in terms of a player's own reward maximization. This optimality, also known as rationality in game theory [52], leads

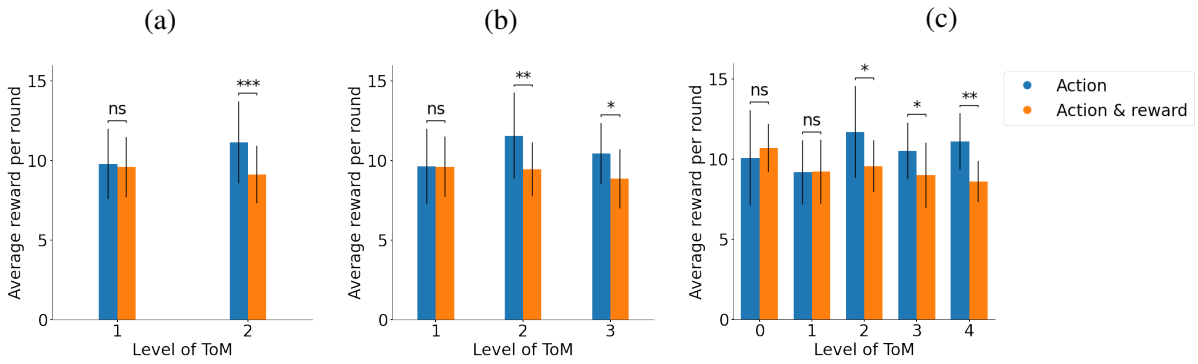


Figure 6.5: **Relationship between levels of ToM, information and reward in the PD task.** Subjects with higher than 1 level of ToM significantly gain more reward in average when only actions of others are available to the player. **a)** The average gained reward of subjects in each round of the game between in the two sessions when using the mixed model of level-1 and level-2 ToM. **b)** same as (a) but level-3 ToM was also included in the mixed model. **c)** same as (a) and (b) but with all levels from 0 to 4. "ns": $p \geq .05$, "*": $p < .05$, "**": $p < .01$, "***": $p < .001$.

to a Nash equilibrium when the depth increases to infinity [55, 69]. We tested this for our public good game, in which all free-rides corresponds to a Nash equilibrium, and also prisoner's dilemma, where all defects is the Nash equilibrium. Specifically, we fit different levels of ToM to the data and calculated the average contribution (in PGG) or cooperation/choosing blue (in PD) rate of the subject predicted by the model.

As seen in Figure 6.6a , the predicted contribution rate decreased to 0 gradually as the level of ToM increased, despite being fit to a dataset with a contribution rate significantly higher than 0. A player whose only goal is maximizing their own reward does not contribute in the last round as there are no future rounds. In fact, consistent with the principle of conformity, the most important effect of contribution is increasing the contribution rate of others to produce more reward in the future. As the level of ToM increases, the player free-rides in earlier rounds because others (mod-

eled as optimal agents) would be expected to free ride in later rounds. Thus, using higher levels of ToM shifts free-riding towards the first round, decreasing the contribution rate over all rounds to 0 (Figure 6.6a).

The prediction about the rate of choosing blue, analogous to cooperation, in the PD experiment follows the same logic. In one-shot prisoner's dilemma the player should defect (here choose yellow) [52]. In a multi-round game this happens in the last round where there is no future, but shifts toward the first round as the level of ToM increases. Therefore, although the model was being fit to a task with significantly higher than zero cooperation rate, the prediction about cooperation rate gradually decreased to 0 with increase of level of ToM (Figure 6.6b).

One of the core concepts of our frameworks is the assumption of "mind of the average group member". On one hand, as exemplified in our first two experiments, tracking individuals are not always possible due to anonymity. Even when each player's action is visible to others, tracking each person separately especially in large groups is computationally very expensive. On the other hand, this assumption is not completely correct as there is a large variance between individuals' internal model and strategies. Here we show that despite of this potential issue, the belief of an average group member actually explains joint beliefs about each individual very accurately.

Our model basically estimates the joint probability of multiple Beta distributions (one for each player) with one Beta distribution. Our simulations, using method of moments [111], show that this estimation is almost always very accurate (figure 6.6c). In fact, for this reason this approach has been also used in machine learning [1]. The estimation does not work only with specific settings where some distributions have a very strong tendency toward one of the options, e.g. $\alpha = 100$, $\beta = 1$ and one of the distributions has no or very little preference over each option, i.e. $\alpha \approx \beta$ (figure 6.6d). Using parameters obtained from the public goods game experiment, we estimated the average of samples generated by $N = 5$ Beta distribution with one single Beta distribution. The largest KL-divergence between the real and estimated distribution (shown in figure 6.6c) was 0.002 and the largest Kl-divergence to entropy of the real distribution ratio was .03%. In figure 6.6d, we also show an example of a case where a single Beta distribution is not an accurate estimate of multiple Beta distributions with KL-divergence of 0.25, which was 3.2% of the entropy of the

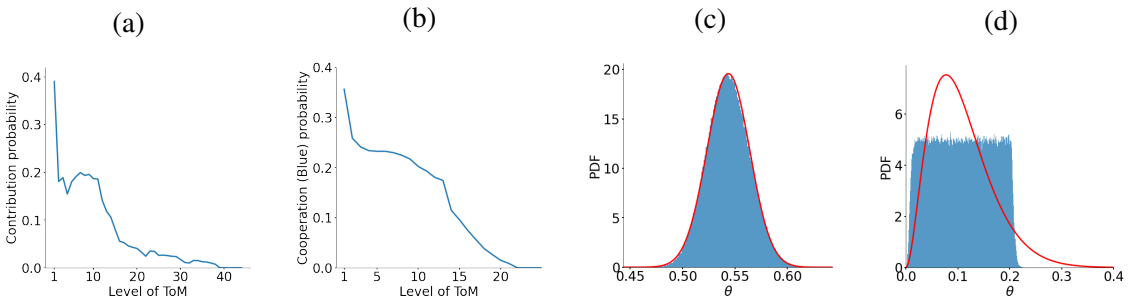


Figure 6.6: **Looking deeper into assumptions of multi-level group decision making framework. a)** In the public good game, as the level of ToM becomes higher, the average contribution rate converges to zero (free riding in all rounds corresponds to a Nash equilibrium). **b)** Similarly in the prisoner's dilemma, as the level of ToM becomes higher, the average cooperation (choosing blue in our experiment) rate converges to zero (defect in all rounds is the Nash equilibrium). **c)** One Beta distribution is an accurate estimate of the joint distribution of multiple Beta distributions. The blue lines show a normalized histogram obtained from average samples of 5 different beta distributions each with parameters randomly picked from our PGG results. The red curve shows the approximated Beta distribution for these samples. This panel shows the worst approximation in terms of KL-divergence from randomly selecting 5 sets of parameter from PGG parameters (α_1 and β_1 pairs) 1000 times. **d)** Inaccurate approximation of multiple Beta distributions by one Beta distribution is possible, but happens in very specific and unrealistic situations involving some distributions with very strong preference over one choice, and some with no preference at all. Specifically, this panel is generated from 4 Beta(1, 100) and 1 Beta(1, 1). Blue lines determine the histogram of average samples and the red curve is the fitted Beta distribution.

actual distribution.

6.5 Discussion

We presented a new Bayesian framework for modeling conformity and multiple levels of theory of mind in collective decision making. To our knowledge, this is the first time that a mathematical model has unified conformity, theory of mind, and utility maximization. Moreover, this framework is the first multi-level ToM model for collective group decision making. Previous models covered either only two-person interactions [165, 4, 36, 69] or only a single level of ToM [140]. Cognitive Hierarchy (CH) modeling [17] is probably the closest work to ours as it also includes multi-level reasoning and could be applied to group decision making. CH modeling however does not model the belief of the subjects through formal and universal principles of Bayesian reasoning. Most importantly, it does not offer a generalizable strategy for level-0 agent [17]. Conformity plays the key role in explaining level-0 and consequently higher levels of ToM in our framework.

We demonstrated the viability of our framework using data from three different experiments, each with different conditions. In addition to quantitative fits to the behavior, levels of ToM that explain subjects' behavior in each experiment were aligned with the properties and conditions of the tasks. In the consensus task, the behavior of the majority of subjects was explained best by a level-0 ToM model. In this task, different choices could be desirable to different subjects but they knew all players had to pick the same choice in order to finish the current game and gain one of the choices. Consistent with this most subjects followed the majority in each game to achieve this goal. Moreover, those who utilize level 1 ToM (according to our framework) achieved their preferred choice (according to the auction) more often as they tried to influence the group to gain more utility.

In the volunteer's dilemma (thresholded PGG) experiment, while the players tried to maximize their own reward, they knew the amount of accumulated reward would be higher if players cooperate with each other. They also knew that more than k contributions would lead to a waste of resources. Overall, strategic game play and reasoning on current and future intentions of others seems more necessary in this task. Therefore, in addition to level-0, many players utilized level-1 and level-2 ToM.

Similar to volunteer's dilemma, keeping the game cooperative was important in the prisoner's dilemma experiment. In fact in this task, as opposed to volunteer's dilemma where a few contributors were enough, every single neighbor's action could change the reward. As a result, keeping others cooperative and strategic game play were more important. Consistent with this intuition, level-1 ToM produced a better fit for nearly all subjects compared to level-0. Also, many subjects utilized level-2, and maybe even level-3 ToM.

We named our framework levels in accordance with previous research on dyadic interactions [165, 69]. For many scientist, especially developmental psychologists and primatologists, theory of mind is reasoning about another agent's hidden intention [130]. This means that theory of mind starts with level 2 of our framework where the agent reasons about others' strategy, especially if we test ToM with the classic study of false belief [151]. On the other hand, our level-1 agent is more sophisticated than level-1 ToM agents in previous frameworks in which they reason about an agent that does not really interact with them, e.g. an agent when random action selection or with a policy independent of actions of others [165, 55]. In fact, this sophistication is actually the consequence of of our level-0 model which practices conformity/imitating the majority. Imitation itself, is in fact, an important social cognitive ability.

Since we used the same set of free parameters for different levels of each task, the main difference between different levels is their cost of computation (here cognitive effort). This cost could not be captured by existing statistical tests for model fitting. There was also a strong correlation between accuracy of different levels for two reasons. First, games with less changes, i.e., consistent selection of an action by the subject, make the fit better for all methods. Second, all levels share the same core model, i.e. conformity, and beside level-0 all tried to maximize the reward of the player. Therefore, we focused on the framework as a whole, and results that are consistent with different criteria for level of ToM selection. Despite of not penalizing the depth of reasoning explicitly, our framework suggested low levels of ToM in our experiments. This was consistent with observed ToM levels of 1 and 2 in fully competitive games in groups according to cognitive hierarchy modeling [17] as well as lack of observance of higher levels even in two-person interaction studies in general [130].

Finally, while we illustrated the approach with only two possible actions, the framework can be easily extended to more actions simply by using multinomial and Dirichlet distributions [102].

6.6 Methods

Higher levels of ToM

If the player uses a common reward function R^o for others in the group, the level-k ToM agent ($k > 1$) is modeled as a POMDP with state space $S = (\alpha, \beta, t)$ where t is the round number ($0 \leq t < H$). The action space remains the same: $A = \{a1, a2\}$. The transition function becomes deterministic as follows. Let the current state be (α, β, t). If $\pi^*_{k-1,t} = a1$ where $\pi_{k-1,t}$ is the policy of the level-$(k-1)$ POMDP with the reward function R^o, the next state is $(\lambda\alpha + N, \lambda\beta, t+1)$ for action $a1$ and $(\lambda\alpha + N - 1, \lambda\beta + 1, t+1)$ for action $a2$. Similarly, if $\pi^*_{k-1,t} = a2$, the next state is $(\lambda\alpha + 1, \lambda\beta + N - 1, t+1)$ for the action of $a1$, and $(\lambda\alpha, \lambda\beta + N, t+1)$ for $a2$. Along the same lines, the reward function is:

$$R((\alpha, \beta, t), a) = R\left((N-1)\mathbb{I}(\pi^*_{k-1,t} = a1), a\right). \tag{6.3}$$

where $\mathbb{I}(x)$ is 1 if event x happens and 0 otherwise. Note that the assumed reward function of others, R^o, could be different from the reward function R of the player but in practice, if there is one reward function for others, it probably applies to all group members including the player.

If the player does not know the reward function of others, they can divide other players into two groups based on their preference (immediate reward) for an action and estimate the reward function of each group separately:

$$R((\alpha, \beta, t), a) = R\left(\frac{(N-1)\alpha}{\alpha + \beta}\mathbb{I}(\pi^{1*}_{k-1,t} = a1) + \frac{(N-1)\beta}{\alpha + \beta}\mathbb{I}(\pi^{2*}_{k-1,t} = a1), a\right). \tag{6.4}$$

Similar to the common R^o reward function case above, we define $\pi^{1*}_{k-1,t}$ and $\pi^{2*}_{k-1,t}$ as policies of level-$(k-1)$ POMDP model and use the reward function R^{o1} with action $a1$ having a higher immediate reward, and R^{o2} with action $a2$ being the more rewarding action.

Model fitting: consensus decision making

By using the first round as the prior, all levels had only one free parameter, the decay rate. It was obtained by a grid search with the precision of .25, i.e., $\lambda \in \{0, .25, .5, .75, 1\}$. For level-1 and higher level models, the reward of obtaining item k was set to the value of that item v_k in the auction. Note that v_k is known.

The possibility of ending up in a terminal state without any reward after the tenth round with 25% chance was incorporated into the transition function to match the reality. For level-2 and higher level models, we divided the other players into two groups with a value of \$2.5 for the favorite choice and \$1.5 for the other choice

$$\begin{cases} R^{o1}(N-1, a1) = 2.5, \quad R^{o1}(0, a2) = 1.5 \\ R^{o2}(N-1, a1) = 1.5, \quad R^{o2}(0, a2) = 2.5 \end{cases} \tag{6.5}$$

For the model-free approach, we set the initial value of each choice to $W + v_k$, where W is a free parameter for each subject. Additionally, we used two parameters to model the time-varying learning rate: $\eta^t = \frac{1}{\kappa_0 + \kappa_1 t}$.

Model fitting: public good game

The fitting process was the same as the previous chapter with one major difference: the 3 free parameters were fit separately for each k. Also, the first round was included in the fits. Similarly, the Q-learning model was the same as Q-learning model in the previous chapter, but with different set of free parameters for each k.

To calculate the average contribution rate at each level, we fitted up to level-30 ToM model to the data. Due to computational limitations, the decay rate was tested with smaller subset of $[.5, .7, .9, .95, .97, .1]$ for levels higher than 3.

Model fitting: prisoner's dilemma

The fitting process was very similar to the PGG's fitting procedure. The only difference was that, actually to be consistent with that procedure, we used horizon of $H = 85$ since the game had 35

rounds more than the games of the PGG experiment.

In the Q-learning model, adding a proportion of reward of others to players' own reward improved the results. Specifically, we changed the reward from r^i (player's own reward for player i) to:

$$r^p + e \sum_{j \in -i} r^j \tag{6.6}$$

Where $-i$ means players other than player i and e is the free parameter.

Moreover, at the beginning of each game, $Q^1(b)$ and $Q^1(a)$ were computed by the assumption that others cooperate(choose blue) with probability of p where p is a free parameter. As in the case of previous tasks, two free parameters were used to describe the learning rate: $\eta^t = \frac{1}{\kappa_0 + \kappa_1 t}$.

To calculate the average contribution rate at each level, we fitted up to level-30 ToM model to the data. Due to computational limitations, the decay rate was tested with smaller subset of $[.7, .9, .95, .97, .1]$ for levels higher than 4.

Chapter 7

CONCLUSIONS

In this thesis we utilize a belief-based decision making framework based on Partially Observable Markov Decision Processes (POMDPs) for modeling decision making in neuroscience and psychology experiments. Due to generalizability of our framework we are able to model a wide range of experiments from perceptual to social decision making tasks. More importantly, since our framework is normative, the quantitative fits of our approach suggest basic principles behind decision making process. The principles include building an internal model of the world, probabilistic reasoning about the current state of the world based on the internal model and observed inputs, and choosing actions that maximizes the total expected reward based on simulating the future.

7.1 Summary of modeling and main contributions

Particularly, we test our framework on perceptual and social decision making experiments due to the key roles of uncertainty and internal model in these experiments. Figure 7.1 gives a summary of our framework applied to each task in previous chapters. We also wish to highlight two important points that we mentioned about using POMDPs for neuroscience studies here. The first point is the existence of a possibly different but plausible state space in the internal model that generates a similar observation distribution compared to the real state space. The second one is efficient policy computation.

As shown in figure 7.1a, in random dots motion discrimination task, frames of moving dots are generated based on a discrete set of coherence levels. The decision maker, however, assumes that coherence levels span a continues wide range because they are not aware of the real-world model. Moreover, these discrete levels produce observations with large overlap. Importantly, this continuous distribution generates very similar observation distribution compared to the real model

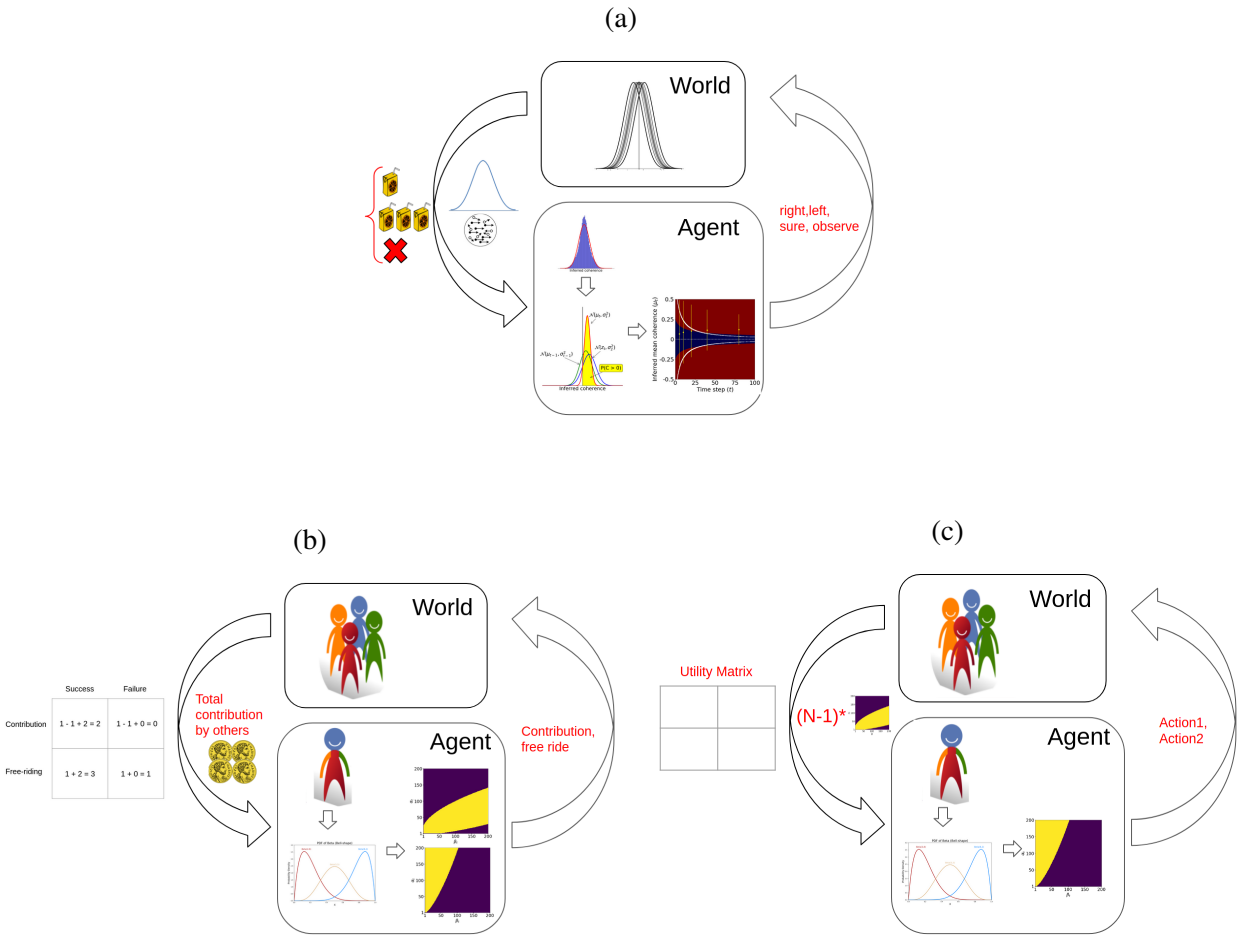


Figure 7.1: **Belief-based decision making framework for different experiments. a)** Model of random dots motion discrimination task with post-decision wagering, presented in chapter 3. **b)** Model of volunteer's dilemma task presented in chapter 5 **c**) Model of level-2 ToM for a collective decision making with two possible choices explained in chapter 6.

as shown in chapter 3. Based on the internal model, the decision maker updates the belief about the motion coherence after each observation.

As the prior and observation functions are Gaussian, the updated belief is always a Gaussian distribution and can be expressed with two parameters of mean and variance. Since the variance

depends on time and some constant variables, the belief can also be expressed with the mean coherence and time. Based on the belief and reward/cost of different options, the decision maker selects one of the actions. Importantly, we showed how one-step look ahead strategy actually selects the optimal action for this task if the cost of observation gathering does not decrease through time. We also explained how the POMDP model can be implemented by an accumulation of evidence to a bound model. These models have been extensively used in explaining the neural data (e.g. [127, 22]).

The belief-based decision making framework also explains observed discrepancies between confidence and accuracy as shown in chapter 4. Particularly, the hard-easy effect, and the opposing effects of the variability of observations on choice and confidence are due to the fact that the decision maker uses an internal model without all information available to the experimenter. The possible difference between sensitivity measurements of accuracy and confidence, as well as dependence of confidence in incorrect trials on the order of confidence report, are because of the role of observation cost. This cost has been largely ignored in common models of perceptual decision making such as SDT and DDM, as they do not model the reward function. Finally, if one ignores the fact that there is a sequence of observations (and not one), and the exact set of used observations are only available to the decision maker, analysis of decision maker's data might lead to a conclusion that choice-congruent observations have a larger weight in determining confidence.

Figure 7.1b demonstrates the volunteer's dilemma modeled by our framework. Here, the actions of others serve as observations of the decision maker. Since actions are anonymous the decision maker assumes that all of them come from an average group member, with a binomial distribution as the likelihood function. We used a Beta distribution to represent the decision maker's belief to make it tractable. With this choice, the belief and consequently the policy could always be expressed with two parameters. The policy map in this task is actually very simple and can be expressed as (at most) three regions: 1) free ride because there are enough volunteers with high probability 2) free ride because there would not be enough volunteer at the moment or in the future 3) contribute to make (or keep) the probability of success high.

A level-2 ToM model of a general group decision making task with two choices is shown in

figure 7.1c. Most of its components are very similar to the components of level-1 ToM model, shown in figure 7.1b. The main difference between the two levels is the observation function. Instead of a binomial distribution, level-2 ToM model assumes that the observations come from level-1 policies of other players. Importantly, in chapter 6 we showed tracking the belief of an average group member is an accurate estimation of tracking joint beliefs of multiple players if we model the beliefs with Beta distribution. In summary our main contributions in this thesis are:

- Modeling choice, performance, belief, and confidence in a unified framework in perceptual decision making
- Presenting a method to obtain confidence from performance in perceptual decision making
- Connecting Bayesian models with accumulation of evidence models
- Explaining common observed discrepancies between perceptual confidence and choice accuracy using an optimal Bayesian framework
- Introducing the concept of "theory of mind of a group" through a Bayesian normative framework tested on a group decision making task
- Building a multi-level theory of mind framework for collective decision making
- Showing that conformity is a basis of collective decision making in groups.

7.2 Future directions

While our framework is practical in presented experiments, we believe it would show its full potential in more complicated tasks. In the field of perceptual decision making, our framework would be more useful in tasks with asymmetric reward and/or prior distribution. In asymmetric cases, sub-optimal decision makers lose more, for example if they do not calculate the prior properly. Unfortunately, even experimental works in this domain are very limited. Another extension could

be in the direction of changing the cost of observation gathering, for example by changing stimulus properties or offering multiple tasks simultaneously to subjects. Cost of observation gathering has not been explored extensively, maybe due to inability of SDT and DDM in modeling it. POMDP, however, provides a very powerful tool to investigate it. These experiments also give us a chance to test our belief-based framework more rigorously.

In the field of social decision making, testing different social games on the same set of subjects could be very beneficial. As shown in chapter 5, we need a lot of data to estimate a subject's level of ToM accurately. Playing the same game over and over, however, is not the solution as the player would learn it and possibly change the strategy and level of ToM over time. Another option is presenting the same game in different sessions to the same subject, but with small modifications in each session.

Combining our computational approach with neuro-imaging is probably the most promising domain yet to be explored in social decision making studies. It would be very interesting to see whether the components of our framework, such as each player's belief about the group, have correlations with the neural activity in specific brain regions. Another interesting direction is using group decision making and our framework in investigating neuropsychiatric diseases.

BIBLIOGRAPHY

[1] Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, pages 33–1. JMLR Workshop and Conference Proceedings, 2012.

[2] M Archetti and I. Scheuring. Coexistence of cooperation and defection in public goods games. *Evolution*, 65(4):1140–1148, 2011.

[3] Hagai Attias. Planning by Probabilistic Inference. In *Proc. of the 9th Int. Workshop on Artificial Intelligence and Statistics*, 2003.

[4] Bahador Bahrami, Karsten Olsen, Peter E Latham, Andreas Roepstorff, Geraint Rees, and Chris D Frith. Optimally interacting minds. *Science*, 329(5995):1081–1085, 2010.

[5] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):0064, 2017.

[6] JD Balakrishnan and Roger Ratcliff. Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance*, 22(3):615, 1996.

[7] S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, and I. Plumb. The "reading the mind in the eyes" test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 42(2):241–251, 2001.

[8] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46, October 1985.

[9] Hannah M Bayer and Paul W Glimcher. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129–141, 2005.

[10] Jeffrey M Beck, Wei Ji Ma, Roozbeh Kiani, Tim Hanks, Anne K Churchland, Jamie Roitman, Michael N Shadlen, Peter E Latham, and Alexandre Pouget. Probabilistic population codes for Bayesian decision making. *Neuron*, 60(6):1142–1152, 2008.

[11] Gordon M Becker, Morris H DeGroot, and Jacob Marschak. Measuring utility by a single-response sequential method. *Behavioral science*, 9(3):226–232, 1964.

[12] Richard Bellman. A markovian decision process. *Journal of Mathematics and Mechanics*, 6(5):679–684, 1957.

[13] Anil Bollimunta and Jochen Ditterich. Local Computation of Decision-Relevant Net Sensory Evidence in Parietal Cortex. *Cerebral Cortex*, 22(4):903–917, April 2012.

[14] Matthew Botvinick, Jane X Wang, Will Dabney, Kevin J Miller, and Zeb Kurth-Nelson. Deep reinforcement learning and its neuroscientific implications. *Neuron*, 2020.

[15] Kenneth H Britten, Michael N Shadlen, William T Newsome, and J Anthony Movshon. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, 12(12):4745–4765, 1992.

[16] George W Brown. Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, 13(1):374–376, 1951.

[17] Colin F Camerer, Teck-Hua Ho, and Juin Kuan Chong. A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences*, 3:157–162, 2015.

[18] Colin F Camerer and Teck Hua Ho. Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874, 1999.

[19] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. In *Advances in Neural Information Processing Systems 32*, pages 5174–5185. Curran Associates, Inc., 2019.

[20] Lucie Charles, Filip Van Opstal, Sébastien Marti, and Stanislas Dehaene. Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage*, 73:80–94, 2013.

[21] Caroline J Charpentier and John P O'Doherty. The application of computational models to social neuroscience: promises and pitfalls. *Social neuroscience*, 13(6):637–647, 2018.

[22] Anne K Churchland, Roozbeh Kiani, and Michael N Shadlen. Decision-making with multiple alternatives. *Nature Neuroscience*, 11(6):693–702, June 2008.

[23] Robert B Cialdini and Melanie R Trost. Social influence: Social norms, conformity and compliance. In *The Handbook of Social Psychology*, volume 2, pages 151–192, 1998.

[24] P. Cisek, G. A. Puskas, and S. El-Murr. Decisions in Changing Conditions: The Urgency-Gating Model. *Journal of Neuroscience*, 29(37):11560–11571, September 2009.

[25] Colin W. Clark and Marc Mangel. The evolutionary advantages of group foraging. *Theoretical Population Biology*, 30(1):45–75, August 1986.

[26] Anne GE Collins and Jeffrey Cockburn. Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, pages 1–11, 2020.

[27] Miguel Costa-Gomes, Vincent P Crawford, and Bruno Broseta. Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5):1193–1235, 2001.

[28] Denis Cousineau et al. Confidence intervals in within-subject designs: A simpler solution to loftus and masson's method. *Tutorials in quantitative methods for psychology*, 1(1):42–45, 2005.

[29] Adam J Culbreth, Andrew Westbrook, Nathaniel D Daw, Matthew Botvinick, and Deanna M Barch. Reduced model-based decision-making in schizophrenia. *Journal of Abnormal Psychology*, 125(6):777, 2016.

[30] Fiery Cushman and Adam Morris. Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, 112(45):13817–13822, 2015.

[31] John M Darley and Bibb Latane. Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, 8(4p1):377, 1968.

[32] Nathaniel D Daw and Peter Dayan. The algorithmic anatomy of model-based evaluation. *Phil. Trans. R. Soc. B*, 369(1655):20130478, 2014.

[33] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

[34] P. Dayan and N. D. Daw. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453, December 2008.

[35] Marie Devaine, Guillaume Hollard, and Jean Daunizeau. Theory of mind: did evolution fool us? *PloS One*, 9(2):e87619, 2014.

[36] Marie Devaine, Guillaume Hollard, and Jean Daunizeau. Theory of Mind: Did Evolution Fool Us? *PLOS ONE*, 9(2):e87619, February 2014.

[37] D. Diekmann. Volunteer's dilemma. *The Journal of Conflict Resolution*, 29(4):605–610, 1985.

[38] Ray J Dolan and Peter Dayan. Goals and habits in the brain. *Neuron*, 80(2):312–325, 2013.

[39] Bradley B Doll, Dylan A Simon, and Nathaniel D Daw. The ubiquity of model-based reinforcement learning. *Current opinion in neurobiology*, 22(6):1075–1081, 2012.

[40] Jan Drugowitsch, Gregory C DeAngelis, Eliana M Klier, Dora E Angelaki, and Alexandre Pouget. Optimal multisensory decision-making in a reaction-time task. *eLife*, 3, June 2014.

[41] Jan Drugowitsch, Ruben Moreno-Bote, Anne K. Churchland, Michael N. Shadlen, and Alexandre Pouget. The Cost of Accumulating Evidence in Perceptual Decision Making. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(11):3612–3628, March 2012.

[42] Jan Drugowitsch, Rubén Moreno-Bote, and Alexandre Pouget. Relation between Belief and Performance in Perceptual Decision Making. *PLoS ONE*, 9(5):e96511, May 2014.

[43] Ernst Fehr and Urs Fischbacher. The nature of human altruism. *Nature*, 425(6960):785, 2003.

[44] Ernst Fehr, Urs Fischbacher, and Simon Gächter. Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, 13(1):1–25, 2002.

[45] Ernst Fehr and Simon Gachter. Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994, 2000.

[46] Christopher R Fetsch, Alexandre Pouget, Gregory C DeAngelis, and Dora E Angelaki. Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*, 15(1):146–154, January 2012.

[47] Christopher R. Fetsch, Roozbeh Kiani, William T. Newsome, and Michael N. Shadlen. Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron*, 83(4):797–804, August 2014.

[48] Stephen M Fleming and Nathaniel D Daw. Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, 124(1):91, 2017.

[49] Stephen M Fleming and Raymond J Dolan. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1338–1349, 2012.

[50] Stephen M Fleming, Elisabeth J Putten, and Nathaniel D Daw. Neural mediators of changes of mind about perceptual decisions. *Nature neuroscience*, page 1, 2018.

[51] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, Giovanni Pezzulo, et al. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68:862–879, 2016.

[52] Drew Fudenberg and Jean Tirole. Game theory, 1991. *Cambridge, Massachusetts*, 393(12):80, 1991.

[53] Samuel J Gershman. What does the free energy principle tell us about the brain? *Neurons, Behavior, Data analysis, and Theory*, 4(1):1–10, 2019.

[54] Paul W Glimcher. *Decisions, uncertainty, and the brain: The science of neuroeconomics*. MIT press, 2004.

[55] P. J. Gmytrasiewicz and P. Doshi. A Framework for Sequential Planning in Multi-Agent Settings. *Journal of Artificial Intelligence Research*, 24:49–79, July 2005.

[56] J. I. Gold and M. N. Shadlen. Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, 5(1):10–16, January 2001.

[57] Joshua I. Gold and Michael N. Shadlen. The Neural Basis of Decision Making. *Annual Review of Neuroscience*, 30(1):535–574, July 2007.

[58] Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems*, pages 766–774, 2010.

[59] Jelena Grujić and Tom Lenaerts. Do people imitate when making decisions? evidence from a spatial prisoner's dilemma experiment. *Royal Society open science*, 7(7):200618, 2020.

[60] Jessica B Hamrick. Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, 29:8–16, 2019.

[61] Jessica B Hamrick. Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, 29:8–16, October 2019.

[62] Balázs Hangya, Joshua I. Sanders, and Adam Kepecs. A Mathematical Framework for Statistical Decision Confidence. *Neural Computation*, 28(9):1840–1858, September 2016.

[63] T. D. Hanks, M. E. Mazurek, R. Kiani, E. Hopp, and M. N. Shadlen. Elapsed Decision Time Affects the Weighting of Prior Probability in a Perceptual Decision Task. *Journal of Neuroscience*, 31(17):6339–6352, April 2011.

[64] Timothy D. Hanks and Christopher Summerfield. Perceptual Decision Making in Rodents, Monkeys, and Humans. *Neuron*, 93(1):15–31, January 2017.

[65] Richard P Heitz. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience*, 8:150, 2014.

[66] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, November 2019.

[67] Y. Huang, A. L. Friesen, T. D. Hanks, M. N. Shadlen, and R. P. N. Rao. How prior probability influences decision making: A unifying probabilistic model. In *Proceedings of The Twenty-sixth Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1277–1285, 2012.

[68] Yanping Huang and Rajesh P. N. Rao. Reward Optimization in the Primate Brain: A Probabilistic Model of Decision Making under Uncertainty. *PLoS ONE*, 8(1):e53344, January 2013.

[69] Andreas Hula, P Read Montague, and Peter Dayan. Monte carlo planning method estimates planning horizons during interactive social exchange. *PLoS Computational Biology*, 11(6):e1004254, 2015.

[70] Quentin J. M. Huys, Tiago V. Maia, and Michael J. Frank. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3):404–413, 2016.

[71] Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, 6(6):1185–1201, 1994.

[72] Allen W. Johnson and Timothy K. Earle. *The Evolution of Human Societies: From Foraging Group to Agrarian State*. Stanford University Press, 2000.

[73] Jessica Joiner, Matthew Piva, Courtney Turrin, and Steve WC Chang. Social learning through prediction error in the brain. *npj Science of Learning*, 2(1):8, 2017.

[74] Peter Juslin, Henrik Olsson, and Mats Bjorkman. Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*, 10(3):189–209, 1997.

[75] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99 – 134, 1998.

[76] Ingmar Kanitscheider, Ruben Coen-Cagli, Adam Kohn, and Alexandre Pouget. Measuring Fisher information accurately in correlated neural populations. *PLOS Computational Biology*, 11(6):e1004218, June 2015.

[77] Rachel L. Kendal, Isabelle Coolen, and Kevin N. Laland. The role of conformity in foraging when personal and social information conflict. *Behavioral Ecology*, 15(2):269–277, March 2004.

[78] Adam Kepecs and Zachary F Mainen. A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1322–1337, May 2012.

[79] Adam Kepecs, Naoshige Uchida, Hatim A. Zariwala, and Zachary F. Mainen. Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455(7210):227–231, September 2008.

[80] Koosha Khalvati, Roozbeh Kiani, and Rajesh PN Rao. Bayesian inference with incomplete knowledge explains perceptual confidence and its deviations from accuracy. *bioRxiv*, 2020.

[81] Koosha Khalvati and Alan K. Mackworth. A fast pairwise heuristic for planning under uncertainty. In *Proceedings of the Twenty-Seventh AAAI conference on Artificial Intelligence*, AAAI'13, pages 503–509, Bellevue, Washington, 2013. AAAI Press.

[82] Koosha Khalvati, Saghar Mirbagheri, Seongmin A Park, Jean-Claude Dreher, and Rajesh PN Rao. A bayesian theory of conformity in collective decision making. In *Advances in Neural Information Processing Systems*, pages 9699–9708, 2019.

[83] Koosha Khalvati, Seongmin A Park, Saghar Mirbagheri, Remi Philippe, Mariateresa Sestito, Jean-Claude Dreher, and Rajesh PN Rao. Modeling other minds: Bayesian inference explains human choices in group decision-making. *Science Advances*, 5(11):eaax8783, 2019.

[84] Koosha Khalvati and Rajesh P Rao. A bayesian framework for modeling confidence in perceptual decision making. In *Advances in Neural Information Processing Systems*, pages 2413–2421, 2015.

[85] R. Kiani, T. D. Hanks, and M. N. Shadlen. Bounded Integration in Parietal Cortex Underlies Decisions Even When Viewing Duration Is Dictated by the Environment. *Journal of Neuroscience*, 28(12):3017–3029, March 2008.

[86] R. Kiani and M. N. Shadlen. Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science*, 324(5928):759–764, May 2009.

[87] Roozbeh Kiani, Leah Corthell, and Michael N. Shadlen. Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, 84(6):1329–1342, December 2014.

[88] Shinichiro Kira, Tianming Yang, and Michael N. Shadlen. A Neural Implementation of Wald's Sequential Probability Ratio Test. *Neuron*, 85(4):861–873, February 2015.

[89] GYULA KovAcs, Rufin Vogels, and Guy A Orban. Cortical correlate of pattern backward masking. *Proceedings of the National Academy of Sciences*, 92(12):5587–5591, 1995.

[90] Armin Lak, Gil M. Costa, Erin Romberg, Alexei A. Koulakov, Zachary F. Mainen, and Adam Kepecs. Orbitofrontal Cortex Is Required for Optimal Waiting Based on Decision Confidence. *Neuron*, 84(1):190–201, October 2014.

[91] Daeyeol Lee. Decision making: from neuroscience to psychiatry. *Neuron*, 78(2):233–248, 2013.

[92] Brian Maniscalco and Hakwan Lau. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1):422–430, March 2012.

[93] Patrick H McAllister. Adaptive approaches to stochastic programming. *Annals of Operations Research*, 30(1):45–62, 1991.

[94] Paul G. Middlebrooks and Marc A. Sommer. Neuronal Correlates of Metacognition in Primate Frontal Cortex. *Neuron*, 75(3):517–530, August 2012.

[95] Kevin J Miller, Matthew M Botvinick, and Carlos D Brody. Dorsal hippocampus contributes to model-based planning. *Nature Neuroscience*, 20(9):1269, 2017.

[96] Ida Momennejad, Evan M Russek, Jin H Cheong, Matthew M Botvinick, ND Daw, and Samuel J Gershman. The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1(9):680, 2017.

[97] Dilip Mookherjee and Barry Sopher. Learning and decision costs in experimental constant sum games. *Games and Economic Behavior*, 19(1):97–132, 1997.

[98] Rubén Moreno-Bote. Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. *Neural computation*, 22(7):1786–1811, 2010.

[99] Michael Moutoussis, Pasco Fearon, Wael El-Deredy, Raymond J Dolan, and Karl J Friston. Bayesian inferences about the self (and others): A review. *Consciousness and cognition*, 25:67–76, 2014.

[100] Michael Moutoussis, Nelson Jesús Trujillo-Barreto, Wael El-Deredy, Raymond Dolan, and Karl Friston. A formal model of interpersonal inference. *Frontiers in human neuroscience*, 8:160, 2014.

[101] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[102] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive computation and machine learning. MIT Press, 2012.

[103] Marnix Naber, Maryam Vaziri Pashkam, and Ken Nakayama. Unintended imitation affects success in a competitive game. *Proceedings of the National Academy of Sciences*, 110(50):20046–20050, 2013.

[104] John A Nevin. Signal detection theory and operant behavior: A review of david m. green and john a. swets' signal detection theory and psychophysics. 1. *Journal of the Experimental Analysis of Behavior*, 12(3):475–480, 1969.

[105] Brian Odegaard, Piercesare Grimaldi, Seong Hah Cho, Megan AK Peters, Hakwan Lau, and Michele A Basso. Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proceedings of the National Academy of Sciences*, page 201711628, 2018.

[106] G. Okazawa, L. Sha, B. A. Purcell, and R. Kiani. Psychophysical reverse correlation reflects both sensory and decision-making processes. *Nat Commun*, 9(1):3479, 08 2018.

[107] Emma C Palmer, Anthony S David, and Stephen M Fleming. Effects of age on metacognitive efficiency. *Consciousness and cognition*, 28:151–160, 2014.

[108] S. A. Park, S. Jeong, and J. Jeong. TV programs that denounce unfair advantage impact women's sensitivity to defection in the public goods game. *Social Neuroscience*, 8, 2013.

[109] Seongmin A Park, Sidney Goïame, David A O'Connor, and Jean-Claude Dreher. Integration of individual and social information for decision-making in groups of different sizes. *PLoS biology*, 15(6):e2001958, 2017.

[110] Seongmin A. Park, Sidney Goïame, David A. O'Connor, and Jean-Claude Dreher. Integration of individual and social information for decision-making in groups of different sizes. *PLOS Biology*, 15(6):e2001958, 2017.

[111] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[112] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[113] Navindra Persaud, Peter McLeod, and Alan Cowey. Post-decision wagering objectively measures awareness. *Nat Neurosci*, 10(2):257–261, February 2007.

[114] Megan A. K. Peters, Thomas Thesen, Yoshiaki D. Ko, Brian Maniscalco, Chad Carlson, Matt Davidson, Werner Doyle, Ruben Kuzniecky, Orrin Devinsky, Eric Halgren, and Hakwan Lau. Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, 1, 2017.

[115] Joelle Pineau, Geoffrey Gordon, and Sebastian Thrun. Anytime point-based approximations for large POMDPs. *Journal of Artificial Intelligence Research*, 27:335–380, 2006.

[116] Alexandre Pouget, Jan Drugowitsch, and Adam Kepecs. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience*, 19(3):366–374, February 2016.

[117] Braden A Purcell and Roozbeh Kiani. Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proceedings of the National Academy of Sciences*, 113(31):E4531–E4540, 2016.

[118] Braden A. Purcell and Roozbeh Kiani. Neural Mechanisms of Post-error Adjustments of Decision Policy in Parietal Cortex. *Neuron*, 89(3):658–671, February 2016.

[119] Dobromir A Rahnev, Brian Maniscalco, Bruce Luber, Hakwan Lau, and Sarah H Lisanby. Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *Journal of neurophysiology*, 107(6):1556–1563, 2011.

[120] R. P. N. Rao. Decision making under uncertainty: a neural model based on partially observable Markov decision processes. *Frontiers in Computational Neuroscience*, 4, 2010.

[121] Rajesh P. N. Rao. Decision Making Under Uncertainty: A Neural Model Based on Partially Observable Markov Decision Processes. *Frontiers in Computational Neuroscience*, 4, 2010.

[122] Roger Ratcliff and Jeffrey N Rouder. Modeling response times for two-choice decisions. *Psychological Science*, 9(5):347–356, 1998.

[123] Roger Ratcliff, Philip L. Smith, Scott D. Brown, and Gail McKoon. Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, 20(4):260–281, April 2016.

[124] BAJ Reddi, Kaleab N Asrress, and Roger HS Carpenter. Accuracy, information, and response time in a saccadic decision task. *Journal of neurophysiology*, 90(5):3538–3546, 2003.

[125] Arbora Resulaj, Roozbeh Kiani, Daniel M Wolpert, and Michael N Shadlen. Changes of mind in decision-making. *Nature*, 461(7261):263, 2009.

[126] James K Rilling and Alan G Sanfey. The neuroscience of social decision-making. *Annual Review of Psychology*, 62:23–48, 2011.

[127] Jamie D. Roitman and Michael N. Shadlen. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 22(21):9475–9489, November 2002.

[128] Stéphane Ross, Joelle Pineau, Sébastien Paquet, and Brahim Chaib-Draa. Online planning algorithms for POMDPs. *Journal of Artificial Intelligence Research*, 32:663–704, 2008.

[129] Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

[130] Tessa Rusch, Saurabh Steixner-Kumar, Prashant Doshi, Michael Spezio, and Jan Gläscher. Theory of mind and decision science: towards a typology of tasks and computational models. *Neuropsychologia*, 146:107488, 2020.

[131] Evan M Russek, Ida Momennejad, Matthew M Botvinick, Samuel J Gershman, and Nathaniel D Daw. Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS computational biology*, 13(9):e1005768, 2017.

[132] C Daniel Salzman and William T Newsome. Neural mechanisms for forming a perceptual decision. *Science*, 264(5156):231–237, 1994.

[133] Joshua I. Sanders, Balázs Hangya, and Adam Kepecs. Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron*, 90(3):499–506, May 2016.

[134] Alan G Sanfey. Social decision-making: insights from game theory and neuroscience. *Science*, 318(5850):598–602, 2007.

[135] Jeffrey D Schall. Neural correlates of decision processes: neural and mental chronometry. *Current Opinion in Neurobiology*, 13(2):182–186, April 2003.

[136] Philipp Schwartenbeck and Karl Friston. Computational phenotyping in psychiatry: a worked example. *eneuro*, 3(4), 2016.

[137] Robert L Seilheimer, Ari Rosenberg, and Dora E Angelaki. Models and processes of multisensory cue combination. *Current Opinion in Neurobiology*, 25:38–46, April 2014.

[138] M N Shadlen and W T Newsome. Motion perception: seeing and deciding. *Proceedings of the National Academy of Sciences of the United States of America*, 93(2):628–633, January 1996.

[139] M. Sherif. *The psychology of social norms*. The psychology of social norms. Harper, Oxford, England, 1936.

[140] Michael Shum, Max Kleiman-Weiner, Michael L. Littman, and Joshua B. Tenenbaum. Theory of minds: Understanding behavior in groups through inverse planning. *CoRR*, abs/1901.06085, 2019.

[141] David Silver and Joel Veness. Monte-carlo planning in large pomdps. In *Advances in Neural Information Processing Systems*, pages 2164–2172, 2010.

[142] E. J. Sondik. The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Operations Research*, 26(2):pp. 282–304, 1978.

[143] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 2018.

[144] Shinsuke Suzuki, Ryo Adachi, Simon Dunne, Peter Bossaerts, and John P O'Doherty. Neural mechanisms underlying human consensus decision-making. *Neuron*, 86(2):591–602, 2015.

[145] Attila Szolnoki and Matjaž Perc. Conformity enhances network reciprocity in evolutionary social dilemmas. *Journal of The Royal Society Interface*, 12(103):20141299, 2015.

[146] Diana I Tamir and Mark A Thornton. Modeling the predictive social mind. *Trends in Cognitive Sciences*, 2018.

[147] Diana I Tamir, Mark A Thornton, Juan Manuel Contreras, and Jason P Mitchell. Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences*, 113(1):194–199, 2016.

[148] Mark A Thornton, Miriam E Weaverdyck, and Diana I Tamir. The social brain automatically predicts others' future mental states. *Journal of Neuroscience*, 39(1):140–148, 2019.

[149] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, Cambridge, MA,, 2005.

[150] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. MIT press, 2005.

[151] Michael Tomasello. How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences*, 115(34):8491–8498, 2018.

[152] John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine learning*, 16(3):185–202, 1994.

[153] Jay J Van Bavel, Katherine Baicker, Paulo S Boggio, Valerio Capraro, Aleksandra Cichocka, Mina Cikara, Molly J Crockett, Alia J Crum, Karen M Douglas, James N Druckman, et al. Using social and behavioural science to support covid-19 pandemic response. *Nature human behaviour*, 4(5):460–471, 2020.

[154] Ronald van Den Berg, Kavitha Anandalingam, Ariel Zylberberg, Roozbeh Kiani, Michael N Shadlen, and Daniel M Wolpert. A common mechanism underlies changes of mind about decisions and confidence. *Elife*, 5:e12192, 2016.

[155] Douglas Vickers. *Decision processes in visual perception*. Academic Press, 1979.

[156] Quang H Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.

[157] Erica van de Waal, Christèle Borgeaud, and Andrew Whiten. Potent Social Learning and Conformity Shape a Wild Primate's Foraging Decisions. *Science*, 340(6131):483–485, April 2013.

[158] Abraham Wald and Jacob Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, pages 326–339, 1948.

[159] Michael L Waskom, Gouki Okazawa, and Roozbeh Kiani. Designing and interpreting psychophysical investigations of cognition. *Neuron*, 104(1):100–112, 2019.

[160] Andrew Whiten, Victoria Horner, and Frans B. M. de Waal. Conformity to cultural norms of tool use in chimpanzees. *Nature*, 437(7059):737, September 2005.

[161] Julian A Wills, Leor Hackel, Oriel FeldmanHall, Philip Pärnamets, and Jay J Van Bavel. The social neuroscience of cooperation. *The Cognitive Neurosciences VI*, 2019.

[162] M. Wunder, S. Suri, and D. J. Watts. Empirical agent based models of cooperation in public goods games. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce (EC)*, pages 891–908, 2013.

[163] Ting Xiang, Debajyoti Ray, Terry Lohrenz, Peter Dayan, and P Read Montague. Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS computational biology*, 8(12):e1002841, 2012.

[164] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[165] Wako Yoshida, Ray J. Dolan, and Karl J. Friston. Game Theory of Mind. *PLOS Computational Biology*, 4(12):e1000254, December 2008.

[166] Wako Yoshida, Ben Seymour, Karl J. Friston, and Raymond J. Dolan. Neural mechanisms of belief inference during cooperative games. *Journal of Neuroscience*, 30(32):10744–10751, 2010.

[167] Jingyang Zhou, Noah C Benson, Kendrick N Kay, and Jonathan Winawer. Compressive temporal summation in human visual cortex. *Journal of Neuroscience*, 38(3):691–709, 2018.

[168] Ariel Zylberberg, Pablo Barttfeld, and Mariano Sigman. The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, 6, September 2012.

[169] Ariel Zylberberg, Christopher R Fetsch, and Michael N Shadlen. The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *eLife*, 5, October 2016.